Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

msii ÉCOLE DOCTORALE

DOCTORAL SCHOOL MSII

ICube Laboratory (UMR 7357)

THESIS presented by :

Idris Hamoud

Defended on: **June 25th 2025**

For obtaining the degree of: **Doctor of Philosophy from the University of Strasbourg**

Field: **Computer Science**

Data-efficient Multimodal Learning by exploiting Scene Semantics for Operating Room Workflow Monitoring

Thesis Directors:

Prof. Nicolas Padoy	Professor, Université de Strasbourg
Prof. Didier Mutter	Professor, HUS, Université de Strasbourg

REVIEWERS:

Prof. Marie-Odile Berger	Research Director, INRIA Nancy Grand Est
Prof. Shlomi Laufer	Assistant Professor, Technion

EXAMINERS:

Dr Vinkle Srivastav	Research Scientist, Université de Strasbourg
Dr Omid Mohareri	Research Manager, Intuitive Surgical Inc.

Abstract

Video recordings of operating room (OR) workflows are invaluable for studying and improving teamwork among clinicians. Automating the recognition of clinical activities in these videos is critical for applications such as modeling interactions and enhancing safety and operational efficiency. However, current methods largely depend on fully supervised training, making datasets even harder to generate and often failing to generalize across ORs with different camera setups. Existing self-supervised techniques focus on appearance-based tasks, overlooking vital semantic information like object detection and human pose data. Incorporating these semantic elements can narrow domain gaps and reduce the need for extensive labeling. This thesis proposes new self-supervised approaches to develop recognition approaches for monitoring OR workflows by emphasizing these "abstract" or semantic modalities. Such modalities are more cost-effective and easier to obtain than manual annotations. Building on recent advancements in self-supervised learning for computer vision, the proposed methods utilize masked autoencoders, multimodal contrastive learning, and carefully designed pretext tasks. Ultimately, this work aims to minimize labeling requirements and bolster the scalability and adaptability of surgical workflow monitoring.

Keywords: Deep learning - Computer Vision - Self-supervised Learning - Multimodal Learning - OR Workflow - Video Understanding

Résumé

L'analyse holistique de vidéos du bloc opératoire est essentielle pour le développement de modèles d'intelligence artificielle capables de reconnaître automatiquement et précisément les différentes étapes du workflow chirurgical. Cette reconnaissance automatique permettrait de créer des modèles d'aide à la décision améliorant la sécurité, l'efficacité et le temps d'utilisation du bloc opératoire. Les méthodes actuelles reposent sur l'apprentissage supervisé, nécessitant beaucoup de données étiquetées, et ne permettant pas un transfert facile dans des blocs operatoires disposant d'un positionnement de camera different. Cette thèse propose de nouvelles approches auto-supervisées pour développer des méthodes d'analyse du déroulement des activités opératoires, en mettant l'accent sur des modalités *abstraites* ou sémantiques telles que la detection d'objet ou l'estimation de pose des cliniciens. S'appuyant sur les avancées récentes de l'apprentissage auto-supervisé en vision par ordinateur, les méthodes proposées utilisent des autoencodeurs masqués, l'apprentissage contrastif multimodal et des tâches prétextuelles soigneusement conçues. L'utilisation de ces modalités moins coûteuses en annotations permettra la mise en place de ces méthodes dans des contextes cliniques réels.

Mots-Clés: Apprentissage profond - Vision par Ordinateur - Apprentissage auto-supervisé - Apprentissage Multimodal - Workflow au Bloc Opératoire - Analyse de Vidéo

Acknowledgements

Although this is 'my' doctoral thesis, its completion would not have been possible without the collaboration, mentorship, and support of several individuals to whom I would like to express my heartfelt gratitude.

This manuscript represents the culmination of several years of study as part of a doctoral program at the CAMMA lab at the University of Strasbourg, conducted in close collaboration with industrial partners from Intuitive Surgical. Before delving into the research, I would like to take a moment to acknowledge and thank the many people who have supported me along this journey.

First and foremost, I would like to thank my jury members: Pr. Marie-Odile Berger and Pr. Shlomi Laufer. It was an honor to have my work reviewed by such esteemed and influential members of the community.

My deepest thanks go to my primary supervisor, Pr. Nicolas Padoy. I began this thesis with only a surface-level understanding of research and modest coding skills. Your guidance gave me the confidence to explore new directions, some successful, others less so, but always instructive. These experiences allowed me to understand the field and its vast possibilities better. I'm also profoundly grateful to Dr. Omid Mohareri and Dr. Muhammad Abdullah Jamal, our industrial collaborators, whose support made this project possible. Thank you for your trust, for granting me access to invaluable datasets and resources, and for welcoming me during my internship in Sunnyvale. Your belief in my potential empowered me to pursue original research with confidence. Thanks also to Pr. Didier Mutter for co-supervising this thesis and granting access to the clinical data for our experiments.

I would also like to thank every member of the CAMMA team, past and present. Your camaraderie and kindness created a warm, welcoming environment I will miss deeply. I'll reflect fondly on our dinners at Massala, the basketball games at Basket Center, and the bouldering sessions. Special thanks to my fellow PhD friends, Saurav, Adit, and Adrien, who have inspired and supported me throughout these years. A particular thank-you goes to Vinkle for his incredible patience and countless hours spent helping me refine my writing.

Beyond academic and professional support, I've been incredibly lucky to be surrounded by wonderful friends these past few years. Thanks to my roommates, especially Colin and Flore, for being there when my motivation was low. I cherish the memories we've created and hope our paths cross again. A special shout-out to my best friend Lucas for coming to visit, that unforgettable paragliding flight, and the countless hours spent debating politics. I'm thankful for all the friends I met along the way, Anaïs, Gaétan, and many more, for the nights out, the music festivals, and all the joy and laughter we shared.

A deeply heartfelt thank-you goes to the woman who has shared my life for more than five years, Léna. Distance has never kept us apart, and I will always cherish our weekends meeting halfway in Paris and the life we are building together. I could never have done this without your unwavering support, warmth, and the sunshine you bring into my life. Thank you for helping me grow into a better version of myself. I love you endlessly.

And of course, this would not be complete without thanking my family. To my grandmother Zaza: your love and life lessons have shaped who I am today. You taught me how to read and helped forge the resilience I carry with me. You'll always have a place in my heart. I know you're watching over me. To my aunt Nezly and uncle Kamel, thank you for stepping in like second parents and always being there for me with love and support. To my brothers, Amir and Adam, I hope to spend more time with you now that this chapter is coming to a close. And finally, to my parents, Mounia and Hamoudi: thank you for your unconditional love and unwavering support. The older I get, the more I admire your strength and character. To my mom, for her fierce drive to protect those she loves, and to my dad, for his calm and steady presence: I love you both more than words can say. I hope that one day I'll be able to return the support you've always given me.

Contents

List of Figures

List of Tables

1

Introduction

Children's games constitute the most admirable social institutions.

Jean Piaget

‘

Contents

1.1 Background

The widespread adoption of Artificial Intelligence (AI) across various industries creates numerous opportunities [Moencks 2022]. Driven by rapid advancements in digital technologies and the generation of vast, diverse datasets [Press 2013], AI continues to transform many aspects of daily life [Allen 2013].

This unprecedented transformation has made everyday tools like chatbots, navigation apps [Cai 2024], and AI-driven search engines widely accessible to the general public, largely due to open availability and user-friendly implementations driven by breakthroughs in models such as OpenAI's ChatGPT [Achiam 2023] and Google's Gemini [Reid 2024]. These innovations were made possible by significant improvements in hardware and the resurgence of deep learning, a branch of machine learning previously considered obsolete. Deep learning models optimize themselves when provided sufficient quality data, which is abundantly available online today. Consequently, these models are now the predominant approach in machine learning. The rapid maturation of AI technologies has allowed their integration into sensitive sectors such as finance and healthcare. AI's powerful data-processing capabilities in finance facilitate accurate predictions of stock market trends influenced by social and political dynamics [Bollen 2011]. Similarly, the sports industry increasingly employs AI-driven analytics, enabling teams to collaborate with data scientists to devise strategies based on historical performance data [Tuyls 2020] (cf. Fig. 1.1). Such analyses address critical factors including injury prevention, fatigue management, team coordination, and personalized coaching informed by psychological player profiles [Tuyls 2020, Wang 2023c].

Inspired by advancements in other sectors, AI shows considerable promise for healthcare. As Organisation for Economic Co-operation and Development (OECD) countries face an aging population, individuals over 60 years of age currently exceeding one billion and projected to reach two billion by mid-century [OECD 2021], healthcare systems are increasingly strained. Public hospitals often struggle with staffing shortages, making it challenging to provide adequate patient care. AI offers potential solutions by improving hospital efficiency and automating routine, time-intensive tasks such as scheduling and administrative inquiries [Topol 2025]. Furthermore, AI-enabled initial patient screenings and triage processes allow healthcare providers to prioritize critical cases, significantly reducing clinician workloads and mitigating systemic bottlenecks.

As reported by a 2018 study [Childers 2018], surgical interventions account for approximately one-third of total healthcare expenditures in the United States. Optimizing surgical practices represents a critical opportunity for improving overall system safety and efficiency. By leveraging operating room video monitoring and human factors analysis, data-driven ap-

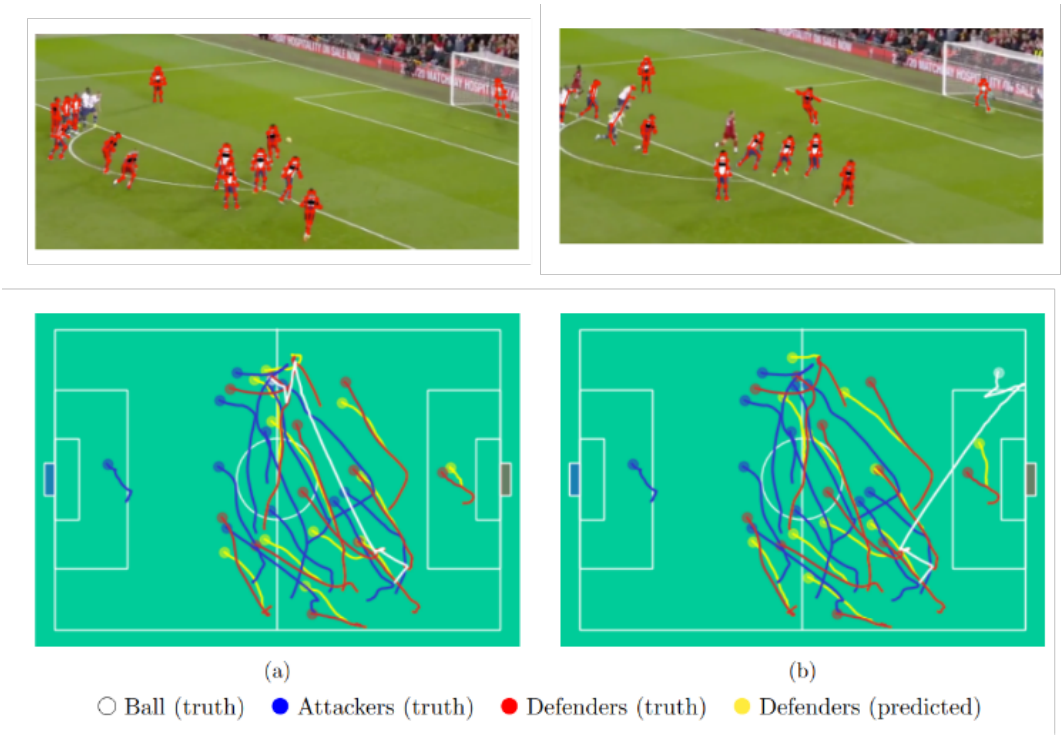

Figure 1.1: Above: Example of a pose estimation model applied to a penalty kick. The pose estimation is acquired using the top-down approach from [Papandreou 2017]. Below: The prediction of defender motion is based on ball and attacking tracking information, which shows different predictions for right and left based on two distinct ball movements. Courtesy of [Tuyls 2020].

plications hold significant promise for enhancing surgical safety, improving procedural outcomes, and streamlining OR management.

Seamless collaboration between surgeons and intelligent systems forms the foundation of AI's impact in the operating room. Unobtrusive visual overlays lighten surgeons' cognitive load, while transparent confidence readouts let them calibrate trust without forfeiting autonomy [Acar 2025, Sakamoto 2024]. Carefully timed, low-distraction alerts further guard against fatigue and preserve focus during critical phases [Fallon 2024]. Together, these design choices pave the way for AI to revolutionize the operating room.

1.2 Revolutionizing Operating Rooms with AI

The operating room is a specialized hospital unit where multidisciplinary healthcare teams collaborate to perform medical procedures using both manual and instrument-assisted techniques. Recent advances in imaging technologies, particularly those driven by nuclear medicine and the de-

velopment of sophisticated surgical instruments, have significantly transformed the modern OR. As illustrated in Fig. A.1, contemporary ORs are highly technological environments. These advancements have also altered surgical methods; for example, procedures like cholecystectomies that once required open surgery are now commonly performed using minimally invasive surgery (MIS). MIS techniques involve small incisions through which surgeons insert specialized instruments and miniature cameras, resulting in reduced blood loss and shorter hospital stays [Bosch 2002].

Robotic-assisted surgery (RAS) extends the principles of MIS by introducing robotic platforms that enhance precision, control, and surgeon ergonomics. These systems provide high-definition, magnified 3D views of the surgical field, allowing for greater instrument articulation, reducing physical strain, and improving operative accuracy.

Modern ORs offer advanced capabilities but involve high operational costs due to the need for specialized staff training and expensive equipment. According to [Childers 2018], the cost of utilizing an OR can reach up to $36 per minute, with even higher expenses incurred during robotic-assisted surgeries due to longer procedures and additional training requirements. Despite these increased costs, the ergonomic improvements provided by modern surgical equipment partly justify the investment [Wee 2020]. Empirical studies have demonstrated that advanced surgical procedures often improve patient outcomes, reduce complication rates, and support quicker recovery times [Reddy 2023].

Multiple case studies have investigated the challenges of effectively integrating robotic-assisted surgery into today's OR environments [Catchpole 2018, Cofran 2021, Catchpole 2024]. Communication issues are among the most significant barriers that can disrupt workflow and extend surgery duration. Effective teamwork, especially during critical phases such as docking surgical robots [Cofran 2021], is essential for maintaining smooth OR operations. Another crucial concern for clinicians and hospital administrators is prolonged turnover times, the intervals between consecutive procedures in the same OR [Souders 2017]. Delays during turnover can significantly disrupt perioperative workflow, making this a critical quality metric [Macario 2006]. Recent advances, such as process standardization and video-based teamwork evaluations, offer promising solutions to address these delays [Rosen 2018]. Analyzing these issues through a socio-technical systems perspective, similar to approaches used in high-risk industries like aviation, can provide valuable insights.

1.2.1 The OR as a Socio-technical System

In this section, we examine how modern ORs, particularly robotic-assisted ORs, can be analyzed through the framework of socio-technical systems [Dias 2020]. Socio-technical theory [Cherns 1976, Baxter 2011] em-

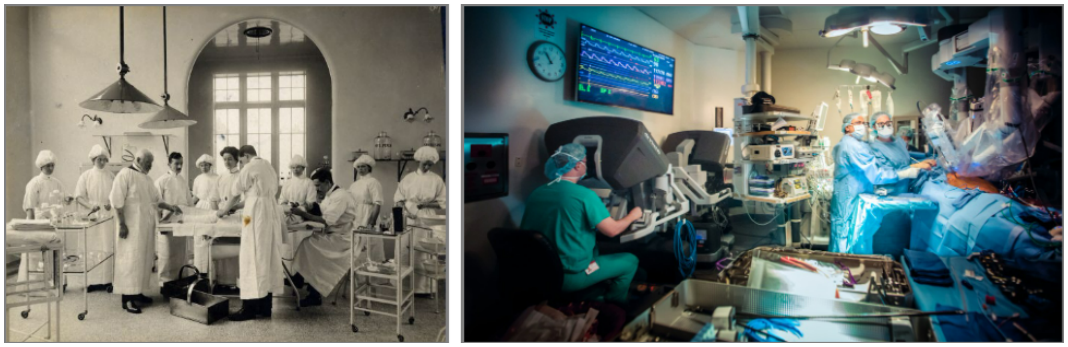

Figure 1.2: This figure shows the evolution of surgical practices from the early 20th century to today. It also illustrates the changes in operating theatres, highlighting the transition to a more modern, highly technological, and cluttered environment. Courtesy of [Lefkowitz 2018]

phasizes the central role of teams, highlighting the specialization of members and their effective interactions as foundations for creating autonomous and efficient groups, rather than relying solely on individual or hierarchical control. These dynamics significantly shape the surgical environment. Operating rooms contain various medical devices and equipment, including robotic systems, which require coordinated operation by professionals from multiple healthcare disciplines, such as surgery, anesthesiology, and nursing. Team members must navigate constraints imposed by procedural types, equipment complexity, and hospital protocols, resulting in a multifaceted and layered system. To manage this complexity, team members collaborate through defined tasks crucial for surgical success, including preparation of equipment, direct patient support, and timely communication regarding tool usage [Anne-Sophie 2009]. Communication within surgical settings is multifaceted; it includes interpersonal exchanges and the interpretation of meaningful data provided by surgical devices, thus supporting surgeons by reducing cognitive load.

Improving communication, standardizing operational procedures, and fostering a learning healthcare environment through innovation can significantly enhance OR efficiency, leading to better patient outcomes and reduced costs [Lee 2019]. Several core activities highlight the inherently collaborative nature of surgical teams. For instance, the seamless exchange of surgical instruments between the scrub nurse and surgeon exemplifies precise teamwork [Korkiakangas 2014, Svensson 2007]. Similarly, clear and effective instruction from surgeons that is promptly followed by the team is critical for procedural success [Svensson 2009]. Within anesthesia teams, dynamic coordination relies on mutual and continuous monitoring among team members [Undre 2006]. Additionally, laparoscopic procedures heavily depend on collective interpretation and identification of surgical land-

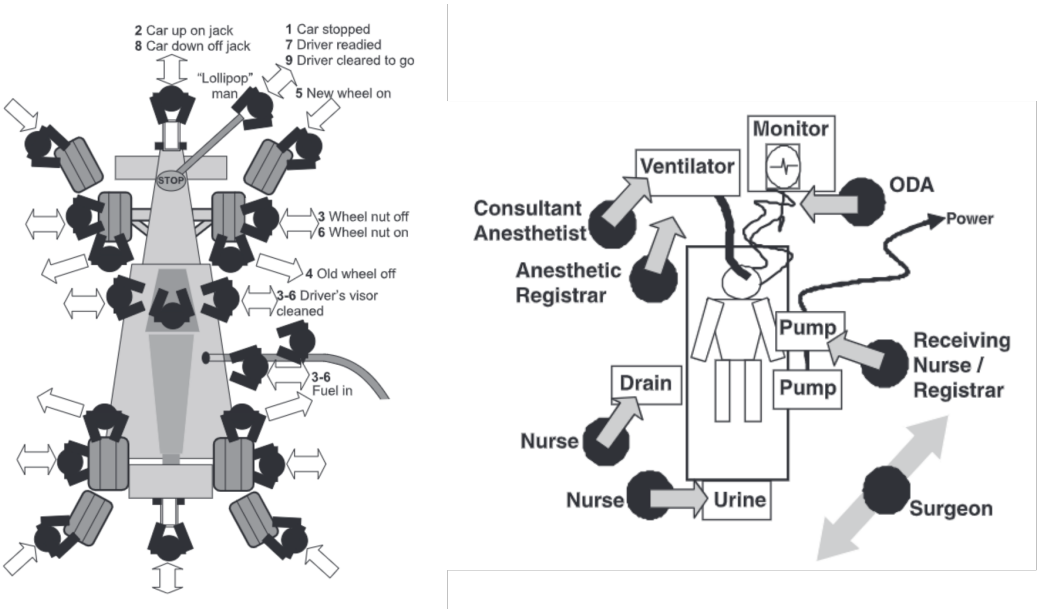

Figure 1.3: Side-by-side comparison of team coordination during pit stops in Formula 1 and during handoffs from the ICU to the OR. ODA stands for Operating Department Assistant, the peri-operative practitioner who works alongside the anaesthetist. Courtesy of [Catchpole 2007]

marks and patient anatomy, further underscoring the importance of collaboration [Koschmann 2011]. Collaboration extends beyond the scope of individual surgeries, which is evident in organizational-level scheduling and time management coordination.

Integrating robotic systems introduces significant shifts in OR dynamics, presenting potential benefits and risks. Several studies [Catchpole 2015, Schiff 2016, Koch 2022, Sheetz KH 2020] highlight critical challenges affecting the safety and effectiveness of RAS, notably issues related to communication breakdowns and coordination difficulties. Research addressing these teamwork-related challenges has found that strategies adapted from high-performance fields such as motor sports pit-stop techniques (see Fig. 1.2.1) effectively reduce technical errors and enhance overall surgical performance [Catchpole 2007].

1.2.2 AI-Enhanced Operating Room Monitoring: Clinical Applications

The modern OR, equipped with diverse monitoring devices and sensors, generates extensive multimodal data. Ceiling-mounted cameras, in particular, capture comprehensive procedure workflows through RGB imaging, depth streams, or low-resolution Time-of-Flight (ToF) videos. These noninvasive data sources offer valuable insights into teamwork dynamics

and could pave the way for AI-driven tools to monitor and enhance workflow coordination with the overall aim of improving surgical safety. In the following, we discuss applications of various workflow monitoring approaches in the OR.

Situation Awareness in the OR Situational understanding, a human factor defined as the ability to perceive, comprehend, and predict future developments within a current situation, is critical for effective decision-making in socio-technical systems [Patrick 2010]. Maintaining shared situation awareness among clinicians in OR teamwork can pose significant challenges. The cognitive load associated with situation awareness could be partially shifted to AI to mitigate this challenge. This can be achieved through the development of situation-aware systems capable of interpreting ongoing activities and delivering appropriate responses within specific contexts. Digitally enhanced ORs, equipped with multi-modal data acquisition methods such as ceiling-mounted cameras, could facilitate the creation of automated tools for activity detection and real-time tracking of surgical progress.

Researchers are developing innovative methods leveraging computer vision and deep learning to enhance OR efficiency [Padoy 2008, Sharghi 2020, Schmidt 2021, Twinanda 2015]. These methods analyze extensive surgical video footage to recognize specific tasks and generate detailed surgical timelines. By delivering real-time or post-operative insights, such technologies empower healthcare professionals to make better-informed decisions, ultimately optimizing patient outcomes. Several foundational components have also been introduced to support OR workflow monitoring. Examples include semantic scene graph prediction, which detects objects and captures interactions between clinicians and their environment [Özsoy 2022, Özsoy 2023], clinician role prediction [Özsoy 2022], and semantic segmentation of robotic and surgical instruments [Li 2020b].

Radiation Safety Monitoring Mobile fluoroscopy devices utilizing intraoperative X-ray radiation have become essential tools in emergency departments and operating theaters, especially in hybrid surgical procedures. Clinicians must wear protective lead-lined gowns to prevent harmful exposure from repeated radiation. Studies have documented significant health risks associated with repeated exposure to intraoperative X-rays [Carinou 2011, Verellen 1999, Koukorava 2014].

Recent research [Rodas 2017, Krebs 2022] (see Fig. 1.4) has introduced simulation models capable of estimating radiation exposure to different body regions, relying on predefined human body models. However, these simulations could be enhanced by incorporating human pose estimation (HPE) techniques to achieve more precise localization of key anatomical joints at the pixel level. This has driven significant research interest in OR workflow monitoring, particularly in developing advanced 2D and 3D HPE

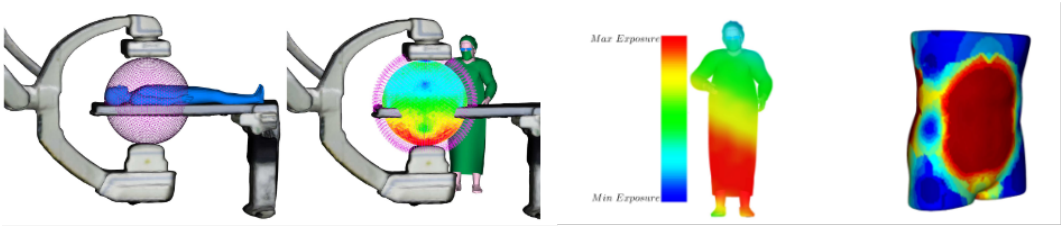

Figure 1.4: Visualization of radiation exposure of staff and patient for a configuration where the X-ray source is under the bed [Rodas 2017]

methods [Srivastav 2018, Srivastav 2021, Srivastav 2020]. Incorporating additional information about surgical activities could help provide detailed reports on total exposure to radiation and per-activity exposure to better optimize radiation.

Monitoring Group Activities in the OR Work in ORs is inherently collaborative and relies heavily on verbal and non-verbal social cues, such as gaze direction and facial expressions. Over time, experienced surgical teams develop specialized communication strategies that enhance their efficiency; for instance, familiarity among team members is known to significantly reduce operative durations [Witmer 2022]. Recent advancements in machine interpretation of non-verbal signals, such as gaze estimation [Nespolo 2022, Gershov 2022] and HPE, present promising opportunities for reducing cognitive workloads during surgical procedures. The utility of such monitoring technologies is particularly pronounced during critical surgical phases [Dias 2020]. Furthermore, structured communication briefings, like Time-Out [Johnston 2009] and StOP? protocols [Keller 2022], have been systematically linked to improved patient safety outcomes [Johnston 2009, Keller 2022]. Automating the detection and differentiation of these briefings could further reinforce safety measures within OR teams [Chen 2025].

Intensive Care Unit Monitoring Similar to ORs, intensive care units (ICUs) require continuous monitoring by clinicians, placing a significant burden on nurses, thereby increasing their workload. Given the abundance of sensors and ceiling-mounted cameras, AI could enable automated monitoring within the ICU. For instance, AI models using HPE could detect potential gestures or facial expressions indicative of pain. In [Davoudi 2019], the authors demonstrated that they could effectively identify delirious patients by employing head pose estimation and limb tracking. Likewise, in [Dai 2024], researchers utilized an HPE-based model to quantify patient movement and assess sedation levels by measuring agitation.

Human–Machine Collaboration in Robotic-Assisted Surgery As discussed in Section 1.2.1, introducing robotic systems significantly alters the dynamics within ORs, bringing both opportunities and chal-

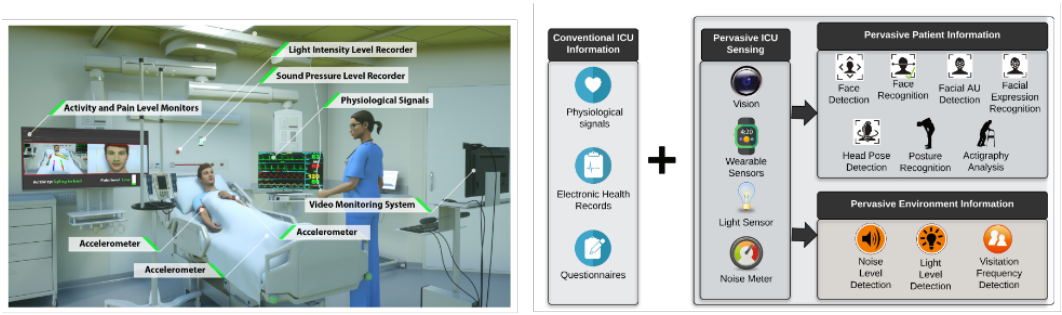

Figure 1.5: Visualization of smart ICU, with the different sensors used to record data such as accelerometer sensors, video monitoring system, light sensor, and sound sensor. Courtesy of [Davoudi 2019]

lenges [Sheetz KH 2020, Cofran 2021]. One promising direction is adaptive autonomy; recent developments in supervised autonomous suturing have demonstrated performance levels comparable to expert surgeons, suggesting that future robotic systems could autonomously manage routine tasks under surgeon supervision [Rivero-Moreno 2024]. Similarly, advancements in robotic scrub-nurse technologies highlight progress toward greater autonomy through enhanced human–machine interfaces (HMIs) and real-time AI monitoring. For example, Wagner et al. [Wagner 2024] have demonstrated that robotic scrub-nurse arms, guided by live laparoscopic video feeds, can accurately anticipate and provide 72% of the surgical instruments required without relying on verbal instructions. This innovation significantly reduces communication demands and enhances team situational awareness. Extending this concept, in [Li 2024], authors introduced RoboNurse-VLA, integrating vision-language-action (VLA) models that combine endoscopic visuals and spoken instructions to grasp and pass previously unseen surgical instruments reliably. This multimodal interaction exemplifies a more natural, intuitive collaboration between robots and surgeons at the task level.

These studies emphasize the critical importance of seamlessly integrating predictive vision, advanced language processing, and ergonomically optimized instrument exchange mechanisms. Such integration is essential for evolving robotic-assisted surgery beyond merely being surgical tools, transforming them into intelligent teammates capable of autonomously sharing and managing surgical tasks.

1.2.3 Challenges for OR Workflow Monitoring Applications

Several challenges complicate using video-based deep learning methods in surgical contexts, including strict data privacy requirements, limited data availability, high labeling costs, and complex, variable operational environments. This change in data distribution could result in a significant failure

of the recognition models. The following sections detail these constraints and explain how they shape our research objectives.

Privacy Preservation Maintaining patient privacy during surgical procedures requires careful management of recorded video content, particularly when images extend beyond the internal views of the patient's body. Such external views can inadvertently disclose sensitive personal information. Various techniques have been developed for endoscopic videos to automatically identify and exclude out-of-body scenes, thus preventing unintended identification of patients or clinical staff [Lavanchy 2023b, Zohar 2020]. In the case of external footage of the OR, the privacy risks are even more pronounced, as these often include visible faces of patients and medical personnel, potentially restricting data collection, which can be mitigated by face detection [Issenhuth 2019] and using low-resolution images during the deployment [Srivastav 2021, Srivastav 2020]. Nevertheless, recording clinician activities remains crucial for comprehensive documentation and analysis of workflows within ORs.

Several privacy-preserving methods have been proposed to mitigate these concerns, including the use of depth cameras [Li 2020a], facial anonymization techniques such as blurring [Flouty 2018, Bastian 2023c], and reducing the resolution of RGB images [Srivastav 2021]. While effective for privacy, these techniques alter the original data distribution compared to conventional vision datasets, adversely affecting the performance of pretrained deep learning models for activity recognition. To counteract this problem, researchers have proposed methods to adapt the models to the privacy-preserving low-resolution images for human pose estimation and person instance segmentation as the downstream tasks [Srivastav 2021, Srivastav 2020]. These approaches have been shown to work well on the downsampled images with a downsampling factor as low as 12x.

Annotation Scarcity Technological advancements in ORs and the incorporation of various sensors, including cameras and endoscopic devices, have significantly increased OR data availability. Such extensive video databases are valuable for developing AI systems to monitor surgical workflows. However, current AI methodologies usually depend on fully supervised approaches, necessitating manual annotation by clinical experts. This reliance poses considerable challenges due to the limited availability of experts and the high costs associated with their specialized knowledge.

To mitigate these annotation demands, an initial strategy involved applying transfer learning by leveraging pretrained feature extractors from existing natural video datasets [Carreira 2017]. Although transfer learning can partially reduce annotation burdens, it might not adequately capture the specific features of surgical data, potentially limiting the effectiveness of the resulting models. To tackle this, researchers have developed unsupervised domain adaptation methods for the task of human pose estimation

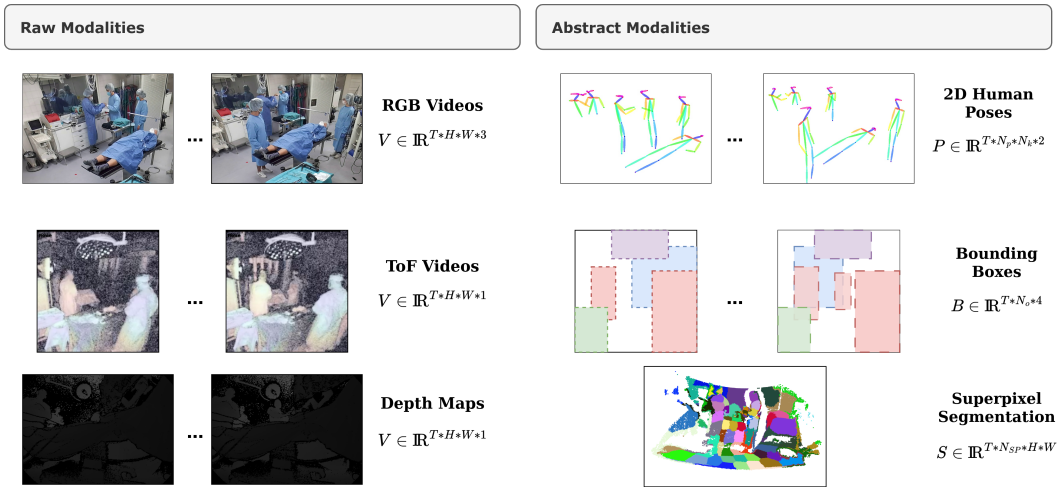

Figure 1.6: In this picture, we present the various modalities in the OR datasets utilized in our thesis. On the left side, we display the raw visual modalities from the datasets: RGB, Time of Flight, and Depth. On the right side, we present abstract modalities such as 2D pose information, object bounding boxes, and superpixel segmentation maps, image courtesy

and person instance segmentation in the OR by adapting models trained on labeled datasets from natural settings to the unlabeled visually distinct OR data [Srivastav 2021, Srivastav 2020].

Self-supervised learning methods are also emerging approaches that circumvent the need for manual labeling by automatically generating training objectives from unlabeled data. These methods enable models to learn features pertinent to the surgical domain directly. Recent progress in self-supervised learning [Chen 2020, Caron 2020, Sun 2022, Wei 2022] has notably enhanced the extraction of meaningful features from unlabeled data, demonstrating promising results across various computer vision applications.

Moreover, data-efficient representation learning methods have emerged to address these challenges by extracting significant features from relatively small datasets, making them especially valuable in medical and surgical contexts. Complementing these approaches, label-efficient learning aims to derive meaningful data representations using minimal or no manual annotations, proving particularly beneficial in environments like ORs, where labeled data is scarce.

Generalization Ability Effective activity recognition in ORs necessitates robust models across various configurations. Contemporary ORs prioritize adaptability to specific patient needs and procedural demands [Satava 2003], resulting in diverse technological setups and camera placements. Consequently, activity recognition systems need to maintain consistent per-

formance despite these environmental differences. Analyzing human motion is crucial in developing reliable OR activity recognition systems, and the interaction between surgical staff and medical devices is another significant factor. Gestures used by surgical personnel to communicate, along with their interactions with medical equipment, present a valuable shared characteristic that can facilitate generalization across different OR environments. Leveraging structured semantic representations derived from these common interactions may effectively address domain gaps. These structured semantic representations, referred to as "abstract modalities" by [Liang 2022], are defined as modalities further removed from direct sensory inputs. As Liang explains:

"Abstract modalities are those farther away from sensors, such as language extracted from speech recordings, objects detected from images, or even abstract concepts like sentiment intensity and object categories."

Therefore, human joint coordinates, object bounding boxes, and object categories serve as abstract modalities that can mitigate domain discrepancies and enhance the generalizability of OR activity recognition systems. Another crucial challenge in monitoring OR workflows is the inherent difficulty in capturing the entire scene using a single camera viewpoint. This is particularly problematic due to significant clinician and instrument movement during critical OR phases, such as patient entry or robotic docking in robot-assisted surgeries. Unlike human observers limited to one viewpoint, AI-based systems can significantly benefit from multi-camera setups offering simultaneous coverage from diverse angles [Kennedy-Metz 2021]. However, practical constraints might limit the availability of multiple cameras or calibrated setups, as described in [Özsoy 2022]. This underscores the importance of leveraging multi-camera streams in a self-supervised approach, enabling the model to infer the OR's spatial arrangement without relying on additional calibration procedures.

1.3 Our Approach

This thesis addresses the annotation bottleneck challenges that hinder operating room video analysis. We propose a self-supervised multimodal pretraining framework that integrates both *raw* (e.g., RGB or Depth video) and *abstract* (e.g., human pose, segmentation masks) modalities. Multimodal pretraining is the process of jointly learning representations from heterogeneous data sources in an unsupervised manner, enabling the model to extract complementary and semantically rich features without relying on manual labels. Our approach harnesses the shared structure across modalities to build robust, generalizable representations, especially suited to surgical scenes' complex, dynamic nature.

Second, our frameworks leverage these abstract modalities, such as pose, as powerful proxies to achieve label-efficient learning. As an abstract

modality, human pose captures essential structural and dynamic characteristics of surgical activities. By exploiting pose-based representations, our methods optimize data utilization and significantly improve labeling efficiency, thus decreasing the dependency on explicit human annotations. This reduction directly addresses practical constraints and privacy concerns associated with data annotation in the OR.

Guided by cognitive science insights into object categorization and the perception of object dynamics [Kahneman 1992, Tenenbaum 2011], we design methods that exploit the motions and actions of objects and clinicians to build compact video representations, thereby enhancing AI performance in practical applications. Our experiments confirm that this principled focus delivers more accurate and efficient video understanding.

Consequently, the core objectives of this thesis are:

(i) Label-efficiency through semantic priors: We leverage inexpensive, unsupervised scene semantic priors to reduce annotation effort while enhancing task accuracy.

(ii) Synergy with self-supervision: Integrating these priors with self-supervised learning objectives yields additional performance gains on downstream benchmarks.

(iii) Viewpoint-robust multimodal pretraining: Our multimodal pretraining demonstrates resilience to viewpoint shifts while retaining complementary information across modalities.

In the following, we briefly describe our contributions.

1.3.1 Analysis of Two Workflow Monitoring Datasets

We provide an in-depth analysis of the two main datasets collected from robotic-assisted surgery operating rooms, which we will use in our experiments. The first dataset, named OR-Seg [Li 2020a], focuses on the semantic segmentation of clinicians and surgical instruments. It consists of images captured by three strategically placed and calibrated Time-of-Flight cameras. The second dataset, OR-AR [Sharghi 2020], which will serve as the central focus of this thesis, also utilizes ToF cameras but employs mobile devices. It represents the most extensive dataset for activity recognition from external cameras in surgical environments. This dataset is particularly valuable due to its diversity, encompassing multiple surgical centers, different procedures, and various surgical teams. Such variability makes it an ideal resource to demonstrate the effectiveness of our self-supervised methods. Additionally, we will assess state-of-the-art (SOTA) object and person detection algorithms using a limited subset of labeled data. Our analysis shows that extracting semantic information from this dataset is possible

with minimal labeling effort, and incorporating these semantic modalities significantly mitigates the annotation bottleneck typically encountered in activity recognition tasks.

1.3.2 Depth-based OR Workflow Monitoring with Superpixel Self-Supervision

As our first contribution, we propose a self-supervised pretext task using superpixel segmentation maps as *abstract* modality. Depth images typically lack texture, making transferring models trained on natural image datasets less effective for operating rooms' semantic segmentation and activity recognition tasks. Nevertheless, as highlighted earlier in paragraph 1.2.3, depth images are beneficial because they preserve patient and clinician privacy while fulfilling the strict OR environment requirements.

To better emphasize depth differences between semantic entities, we introduce a pretext task centered on predicting the average distance between the centroids of projected superpixel clusters, leveraging depth data and camera intrinsic parameters. We hypothesize that depth variations effectively highlight boundaries between semantic regions, thus improving semantic segmentation performance. Additionally, encoding precise distance information between superpixel clusters can enhance downstream activity recognition tasks.

We first employ SLIC [Achanta 2012], a standard superpixel segmentation algorithm, to define superpixel clusters on depth maps. Subsequently, we sample pairs of clusters across images within each dataset and utilize a specialized pooling operator to incorporate geometric knowledge into our embedding space. We demonstrate the effectiveness of our approach on both OR-Seg [Li 2020a] and OR-AR [Sharghi 2020], using data-efficiency protocols where models progressively receive increasing amounts of labeled data in a semi-supervised manner. Our method substantially improves compared to other self-supervised pretraining strategies on both datasets.

1.3.3 Self-Supervised Masked Object Embedding Prediction for Object-Centric OR Activity Recognition

In our second contribution, we introduce a novel activity recognition task based on object layouts, leveraging specialized object and person detectors trained with a minimal subset of annotated bounding boxes from the OR-AR dataset. We propose an MLP-based architecture, called *ST(OR)*^{2}, that temporally integrates bounding box coordinates and object category information to classify video clips according to the depicted activity accurately. We highlight the method's sample-efficient nature through experiments focused on data efficiency.

Further, we extend this method by incorporating a refined architecture utilizing a general-purpose transformer model [Vaswani 2017], taking ad-

vantage of its inherent flexibility in modeling relationships among input elements regardless of positional constraints. Leveraging the transformer's attention mechanism, we propose a self-supervised masked object embedding task, inspired by prior computer vision research [Wei 2022, Sun 2022]. Our results demonstrate that this masked embedding task significantly enhances the performance of our model in recognizing operating room activities within the OR-AR dataset.

1.3.4 Robust OR Activity Recognition via Self-Supervised Multimodal Feature Alignment of Video and Pose Across Multiple Views

Our initial contributions introduced methods that independently leveraged video streams from various perspectives. However, as highlighted in paragraph 1.2.3, integrating multi-camera viewpoints can significantly enhance the recognition of activities within the operating room. In this work, we aim to address the following questions: How can features extracted from different camera viewpoints be aligned effectively in a self-supervised manner to improve our AI model's robustness? Can human pose estimation serve as an intermediary modality to bridge these differing viewpoints?

As our final contribution, we propose a novel self-supervised task, called *PreViPS*, specifically designed for multiview video-pose pretraining. First, we introduce an innovative pose-based action recognition architecture employing a discrete, vector-quantized representation inspired by recent research [Geng 2023]. Next, taking cues from video-language models [Radford 2021, Goel 2022], we develop a two-stream encoder architecture that independently processes pose and RGB data. These streams are then aligned using geometric constraints and a masked pose prediction objective. We demonstrate that our pretraining method enhances performance significantly, both in pose-based and RGB-based activity recognition tasks. Moreover, our pretraining strategy results in performance gains in single-view and cross-view scenarios, even when training and testing viewpoints are entirely distinct. Evaluations on two benchmark datasets, OR-AR [Sharghi 2020] and 4D-OR [Özsoy 2022], underscore our approach's effectiveness. Notably, the marked improvement in pose-only activity recognition supports the utility of our method for privacy-preserving applications, as pose data inherently avoids identifiable information about clinicians and patients.

1.4 Thesis Outline

This thesis investigates the complexities involved in workflow monitoring within operating room environments, particularly focusing on overcoming the labeling bottleneck. We propose multi-modal learning strategies

to enhance annotation efficiency that leverage self-supervision across abstract representations and raw visual data. Our methods incorporate cost-effective, unsupervised prior information, including superpixel segmentation, minimal bounding-box annotations for object detection, and advanced 2D pose estimation techniques for clinicians. We show that integrating these inexpensive annotations significantly boosts labeling efficiency. Moreover, combining these approaches with self-supervised learning objectives further enhances performance in downstream tasks.

The thesis structure is organized as follows:

- Chapter 2 provides an overview of related research in video analysis within computer vision. It begins by discussing unstructured representations and then moves to structured representations using abstract modalities. This chapter also reviews both unimodal and multimodal self-supervised learning methods, highlighting their relevant applications in Surgical Data Science (SDS).

- Chapter 3 describes the OR-AR [Li 2020a], OR-Det [Hamoud 2023], and OR-Seg [Sharghi 2020] datasets, providing comprehensive statistics and benchmarking results for object detection tasks; these datasets form the core of our experimental analysis in subsequent chapters.

- Chapter 4, published in [Hamoud 2022], proposes depth-guided geometric arrangement of superpixel clusters as a self-supervised pretext task, demonstrating improved semantic segmentation and activity recognition.

  HAMOUD, I., KARARGYRIS, A., SHARGHI, A., MOHARERI, O., PADOY, N. (2022). SELF-SUPERVISED LEARNING VIA CLUSTER DISTANCE PREDICTION FOR OPERATING ROOM CONTEXT AWARENESS. IPCAI-INTERNATIONAL JOURNAL OF COMPUTER ASSISTED RADIOLOGY AND SURGERY, 17(8), 1469-1476.

- Chapter 5, published in [Hamoud 2023], presents an object-centric approach for video action recognition, emphasizing data efficiency and integration potential with RGB features. This work has been further enhanced through a transformer-based architecture, enabling a masked object embedding self-supervised objective.

  HAMOUD, I., JAMAL M.A, SRIVASTAV, V., MUTTER, D., PADOY, N., MOHARERI, O. (2023). ST(OR)2: SPATIO-TEMPORAL OBJECT LEVEL REASONING FOR ACTIVITY RECOGNITION IN THE OPERATING ROOM. MEDICAL IMAGING WITH DEEP LEARNING (MIDL)

- Chapter 6, based on [Hamoud 2025], investigates multimodal, multi-view approaches to pose-based action recognition, enhancing portability and multimodal alignment. The work is a journal submission under review.

  HAMOUD, I., SRIVASTAV, V., JAMAL, M. A., MUTTER, D., MOHARERI, O., PADOY, N. (2025). MULTI-VIEW VIDEO-POSE PRETRAINING FOR OPERATING ROOM SURGICAL ACTIVITY RECOGNITION. ARXIV PREPRINT ARXIV:2502.13883.

- The final part, Chapter 7, discusses practical applications and future directions for advancing research in OR surgical activity recognition.

2

Related Work

The highest activity a human being can attain is learning for understanding because to understand is to be free.

Spinoza

Contents

This chapter presents related literature on different methods and concepts relevant to video representation learning, particularly emphasizing their applications in surgical data science. The first section examines global appearance-based models for surgical video action recognition. It traces the evolution from traditional machine learning techniques to advanced deep learning architectures and highlights their practical implications within Surgical Data Science (SDS). The second section focuses specifically on structured semantic representations derived from abstract modalities. As introduced previously (see Section 1.2.3), these modalities, including object detection, human pose estimation, and superpixel segmentation, offer rich and structured semantic information that complements raw visual data. The third section reviews recent advancements in self-supervised learning methodologies for surgical video analysis, exploring both unimodal and multimodal approaches that address the challenges associated with annotated data scarcity. The final section then synthesizes operating-room workflow-monitoring techniques, spanning human-pose estimation to the automated detection of communication protocols, derived solely from external camera views.

2.1 Global Video Representations

This section will discuss global representations that utilize only the *raw* visual modalities. These methods directly leverage raw pixel information to extract a unified representation for a specific frame or video clip.

2.1.1 Traditional Machine Learning Approaches for Video Analysis

Early approaches in video action recognition within computer vision predominantly relied on handcrafted features, broadly categorized into *spatial features* and *spatio-temporal features*. Spatial features characterize individual images using attributes like color [Jain 1996] and texture [Manjunath 1996], typically represented by local descriptors capturing salient spectral information. Sophisticated handcrafted descriptors, such as Scale-Invariant Feature Transform (SIFT) [Lowe 2004] and Speeded-Up Robust Features (SURF) [Bay 2008], were developed to enhance robustness against image variations and augmentations.

Beyond single-frame analysis, *spatio-temporal features* incorporate temporal dynamics within video sequences. Optical flow-based descriptors encode pixel-level motion patterns across frames, such as histograms of oriented optical flow [Chaudhry 2009]. Additionally, three-dimensional occupancy grids have been proposed to merge spatial information from multiple viewpoints, effectively recognizing human actions [Weinland 2007]. Laptev et al. [Laptev 2003] extended the concept of spatial interest points into the temporal dimension, introducing space-time interest points (STIPs) that ro-

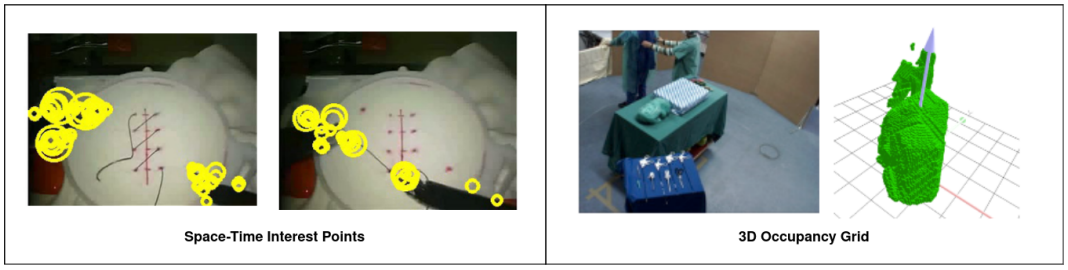

Figure 2.1: Examples of global feature extractors used in early surgical activity recognition methods. Left: Space-Time interest points [Laptev 2003] extracted from video of surgemes [Zappella 2013]. Right: Illustration of the 3D occupancy grids used to extract features for OR workflow activities classification. Courtesy of [Zappella 2013, Padoy 2008]

bustly detect distinct motion patterns, even under challenging conditions such as occlusions and dynamic backgrounds.

Traditional methods often employed dynamic probabilistic graphical models to better model longer-term temporal dependencies, notably Hidden Markov Models (HMMs) [Rabiner 2007]. HMMs effectively represent complex actions as sequences of transitions between hidden states, processing sequences of feature vectors frame by frame. Action recognition in this context involves evaluating the likelihood of observed image sequences fitting a previously learned model.

Applications in SDS Traditional machine learning techniques were extensively applied in early SDS research, combining handcrafted features and probabilistic models. Lalys et al. [Lalys 2010] successfully integrated spatial descriptors (color, texture, shape) with HMMs and Support Vector Machines (SVMs) for surgical phase classification in pituitary surgeries. Similarly, Haro et al. [Haro 2012] leveraged STIPs described by histograms of oriented gradients and oriented optical flow, classifying surgical gestures (surgemes) with SVMs. In parallel, [Padoy 2009] utilized 3D occupancy grids combined with histograms of motion orientations for accurate workflow activity recognition in operating rooms (see Fig. 2.1). Despite their effectiveness, these traditional methods eventually faced limitations, such as challenges in handling complex visual variations and difficulties in automatically learning robust, high-level features from extensive datasets. These pioneering methods established foundational approaches later complemented by deep learning techniques due to their superior scalability and feature representation capabilities.

2.1.2 Deep Learning Approaches for Video Analysis

With handcrafted features and graphical models' limitations in capturing high-level abstractions and complex temporal patterns, deep learning meth-

ods emerged as a powerful alternative for video analysis. These approaches enable end-to-end learning of spatial and temporal features directly from raw data, allowing for greater generalization and adaptability across complex video understanding tasks.

Convolutional Approaches

Initial deep learning models for video analysis extended convolutional neural networks (CNNs) to handle temporal dynamics. Baccouche et al. [Baccouche 2011] introduced one of the earliest approaches, combining 3D convolutions with recurrent layers (LSTMs) to jointly model spatial and temporal dependencies. This architecture was refined with unsupervised strategies for spatio-temporal representation learning [Baccouche 2012], laying the groundwork for subsequent developments.

Modern convolutional approaches fall into two main categories:

- **3D Convolutional Networks:** These networks extend 2D convolutional kernels into the temporal domain, enabling the extraction of motion features across multiple frames [Taylor 2010, Tran 2014, Karpathy 2014]. While effective at capturing fine-grained motion patterns, they introduce significant computational demands and require large datasets for training.

- **Two-stream Architectures:** Proposed by Simonyan et al. [Simonyan 2014], these models process RGB frames and optical flow in separate CNN streams. By learning complementary spatial and motion cues, they achieve improved recognition of actions with varying temporal characteristics. However, reliance on precomputed optical flow increases computational overhead.

Recurrent Neural Networks

Many works combine CNNs with recurrent neural networks (RNNs) to capture sequential dependencies more explicitly. Traditional RNNs, while theoretically suited for sequence modeling, suffer from vanishing gradient problems. Long Short-Term Memory (LSTM) networks address these limitations via gating mechanisms that control information flow [Donahue 2014]. A common pipeline extracts spatial features using a shared CNN and feeds them to LSTMs to model temporal patterns. Despite their effectiveness, these approaches often struggle to model detailed motion due to their reliance on aggregated frame-level features.

Attention-based Approaches

Transformers have recently redefined video analysis by leveraging attention mechanisms that model long-range dependencies without recurrent or convolutional structures. Inspired by their success in natural language processing, Vision Transformers (ViT) [Dosovitskiy 2020] and their video-specific

variants like ViViT [Arnab 2021] apply self-attention to visual tokens, enabling flexible and scalable modeling of spatio-temporal information. However, the quadratic complexity of attention demands architectural optimizations, especially for long video sequences typical in surveillance, education, or medical contexts.

These deep learning frameworks form the foundation of modern video understanding, offering more robust, scalable, and semantically rich representations than traditional handcrafted methods. Their flexibility and performance have made them central to numerous domains, including healthcare and human activity analysis.

Applications in SDS

The adoption of deep learning in SDS has closely mirrored general advances in video analysis. Early approaches leveraged pretrained CNN models such as AlexNet [Krizhevsky 2012]. Twinanda et al. [Twinanda 2016] introduced EndoNet, which extracted deep features using AlexNet and used a combination of SVM and HMM to perform surgical phase recognition on the Cholec80 dataset, a benchmark still widely used today. Their approach reflects early integration of convolutional features with probabilistic temporal models (see Section 2.1.1).

Subsequently, deeper and more specialized networks were explored. Zisimopoulos et al. [Zisimopoulos 2018] proposed DeepPhase, which combined CNNs and LSTMs to model both spatial and temporal cues for phase recognition, marking one of the earliest uses of recurrent models in SDS. This architecture was further refined in EndoLSTM [Twinanda 2017], emphasizing temporal consistency across endoscopic frames.

To improve temporal resolution and handle real-time constraints, Czempiel et al. adapted the MS-TCN framework [Farha 2019] into TeCNO [Czempiel 2020], using causal convolutions to avoid future frame leakage. Recent work by Czempiel et al. [Czempiel 2021] and Gao et al. [Gao 2021] explored transformer-based models with temporal attention. OperA [Czempiel 2021] introduced an attention regularization loss to improve focus on temporally ambiguous frames, while Trans-SVNet [Gao 2021] fused ResNet-derived spatial features with temporal encodings, accelerating inference through parallel processing.

Given the high data requirements of transformers [Dosovitskiy 2020], Bati et al. proposed EndoViT [Batić 2024], employing domain-specific self-supervised pretraining. Liu et al. [Liu 2023] introduced LoViT, which incorporated phase transition maps to better capture surgical phase boundaries. These recent methods highlight adaptations of transformer architectures to the unique constraints of SDS, including long sequences, limited annotated data, and task-specific transitions.

Finally, several studies extended video understanding to the external camera view. Sharghi et al. [Sharghi 2020] employed I3D models for work-

flow recognition from external RGB streams. Schmidt et al. [Schmidt 2021] enhanced this pipeline with cross-view attention to fuse multi-camera perspectives. Both works leveraged GRU networks [Chung 2014] to capture temporal patterns across video clips, reflecting growing interest in contextual, holistic OR activity analysis.

2.2 Structured Video Representations

While global video representations (Section 2.1) primarily rely on raw pixel intensities and aggregated spatio-temporal cues, structured video representations introduce a more semantically organized view of scene dynamics. Instead of encoding a video holistically, these approaches focus on semantically meaningful entities such as objects, body joints, or pixel groupings, capturing their temporal coherence and interactions over time. This perspective draws inspiration from cognitive science. For instance, studies have shown that humans, especially infants, rely on motion and boundary-based cues rather than static appearance when perceiving objects [Scholl 2007]. Similarly, Johansson's seminal work [Johansson 1973] demonstrated that sparse joint motion is sufficient to infer human activities like walking or dancing. Abstracting raw footage into structured components in video understanding provides more interpretable and robust representations, especially in cluttered or occluded environments like operating rooms.

2.2.1 *Abstract* Modalities employed in Video Analysis

Object Presence and Detection transforms raw video into higher-level information by identifying and localizing objects using bounding boxes and category labels [Wang 2018, Materzynska 2019, Herzig 2022, Radevski 2021]. This abstraction enables models to focus on semantically meaningful elements, improving robustness to lighting variations, occlusion, and background noise. By tracking objects' spatial layout and temporal interactions, these approaches support reasoning over the dynamics of objects in a scene.

Human Pose Estimation encodes human body configurations through skeletal keypoints. These representations reduce high-dimensional video frames into structured motion descriptors invariant to appearance and can be analyzed for action recognition. Pose estimation is particularly useful in cluttered settings where body silhouettes are partially occluded, such as surgical environments where medical staff interact closely.

Unsupervised Pixel Grouping refers to over-segmenting an image into superpixels or temporally stable segments. These groupings preserve local visual coherence and can serve as mid-level primitives for learning actions or scene structure without strong supervision. In [Ke 2010], superpixel volumes are compared to action templates via optical flow and vol-

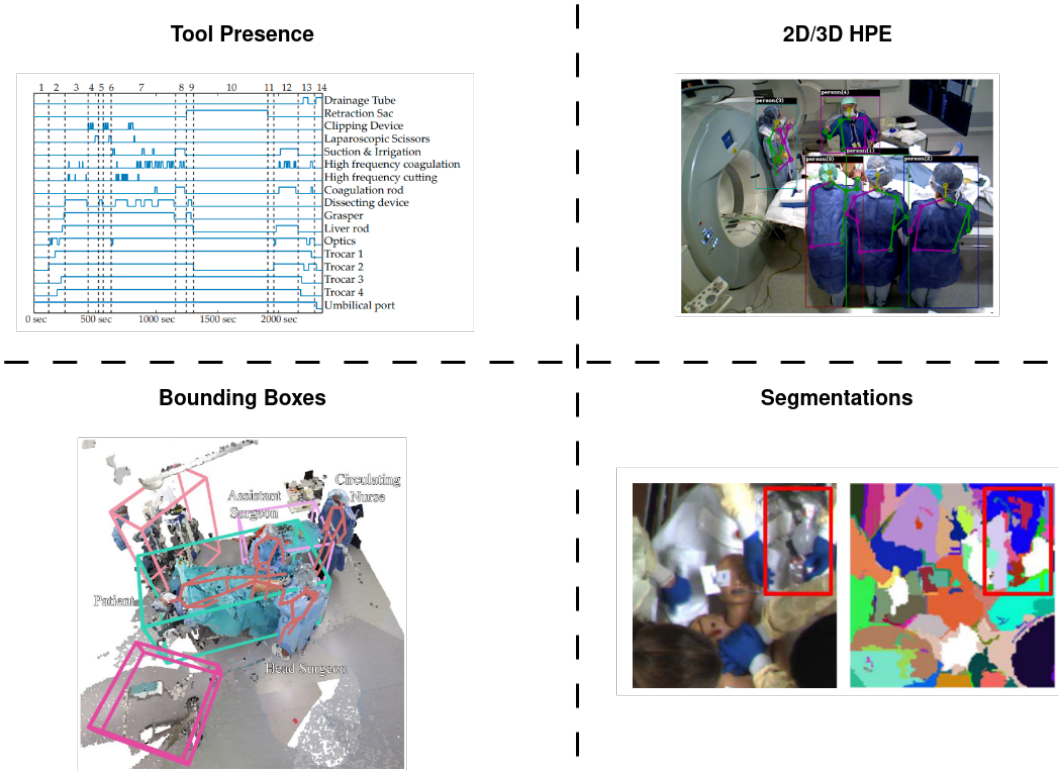

Figure 2.2: Examples of *abstract* modalities used in SDS. Tool presence information [Blum 2008], human pose estimation [Srivastav 2018], object detections [Özsoy 2022] and superpixel segmentation [Chakraborty 2013] using the method developed in [Felzenszwalb 2004]

umetric matching, demonstrating their potential in action recognition (see Figure 2.2).

Object-Centric Approaches for Video Analysis

While global feature models often capture background biases [Carreira 2017, Soomro 2012, Kuehne 2011], object-centric approaches introduce more targeted representations by modeling interactions among detected objects. Early strategies either aggregated object detection outputs globally [Wu 2016] or used graph-based methods such as STRG [Wang 2018] to reason over object relationships across time.

Building on this, STIN [Materzynska 2019] introduced a dual-stream framework using bounding-box and category information, while ORViT [Herzig 2022] enhanced cross-stream communication through attention. ObjectViViT [Zhou 2023] further improved efficiency with strategic token sampling for scalable reasoning.

Applications in SDS Early studies in surgical phase recognition [Padoy 2012, Blum 2008, Padoy 2008] primarily utilized Hidden

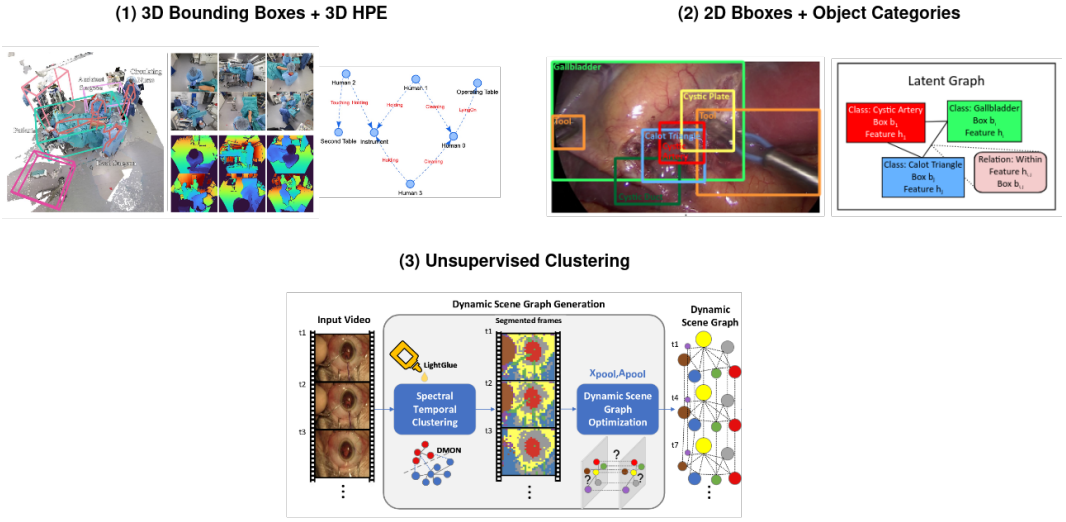

Figure 2.3: Three recent SDS methods using abstract, semantically rich modalities as intermediate representations: (1) role prediction in the OR [Özsoy 2022], (2) critical-view-of-safety assessment [Murali 2023], and (3) phase estimation on CATARACTS [cCaughan Koksal 2024, Hajj 2019].

Markov Models (HMMs) based on surgical tool usage patterns extracted from manually annotated recordings. Despite their effectiveness, these approaches assumed a strictly sequential workflow, limiting their flexibility in handling deviations.

Recent advancements include a graph-based representation proposed in [Murali 2022], employing varying supervision levels from bounding boxes to segmentation maps. This compact latent representation significantly improved performance in assessing the critical view of safety (CVS). Further extending this method, researchers in [Satyanaik 2024] demonstrated its potential for domain adaptation across different clinical settings.

HPE-based Approaches for Video Analysis

Initially, skeleton-based action recognition methods predominantly employed recurrent neural networks and Long Short-Term Memory (LSTM) models [Du 2015], benefiting from the sequential structure of skeletal data. Subsequent developments leveraged convolutional neural networks by converting joint data into 2D heatmaps. The "Ske2Grid" approach [Cai 2023], for instance, transformed skeletal sequences into grids suitable for convolutional operations via bijective upsampling and coordinate mapping.

Alternatively, several studies [Cheng , Cheng 2020, Chen 2021a, Chen 2021b] have modeled skeletal structures as graphs, using GCNs [Kipf 2016] to analyze joint relationships. Although GCNs effectively represent static topologies, their fixed kernels limit adaptability to

variations in skeletal structure.

More recently, transformer-based models [Do 2024, Shi 2020, Duan 2023, Gao 2022] have demonstrated superior capabilities in capturing long-term dependencies between skeletal joints compared to GCN-based methods. Nonetheless, transformers typically require extensive annotation data for training.

Applications in SDS The novel dataset introduced in [Özsoy 2022, Özsoy 2023] includes detailed mock recordings of knee replacement procedures and establishes a pioneering benchmark for semantic scene graph analysis in surgical workflow monitoring. Enriched with multimodal annotations and clinical role predictions, this dataset underscores the significance of accurately capturing object interactions within surgical settings. Additionally, the 2D pose-based methodology presented in [Chen 2025] addresses activity detection tasks such as "Time-out" and "StOP" identification [Keller 2022] within the operating room context.

2.2.2 Pixel Grouping for Structure-Aware Reasoning

The study presented in SANGRIA [cCaughan Koksal 2024], the authors propose an approach beyond simply utilizing labeled spatial information to ground their graph-based representation. They employ a simple spectral clustering technique to generate coarse segmentations, using pretrained DINO [Caron 2021] weights. Each segmented instance is temporally connected using the LightGlue [Lindenberger 2023] feature matcher. The authors reason over the resulting graph using Graph Convolutional Network (GCN) [Kipf 2016] modules and demonstrate competitive results on the CATARACTS [Hajj 2019] benchmarks for surgical phase recognition.

2.3 Self-Supervised Video Representation Learning

Alternative methods utilizing intrinsic information within visual data have emerged to address the data-labeling bottleneck associated with supervised deep learning. Self-supervised learning (SSL), a subset of unsupervised learning, focuses on deriving discriminative features directly from unlabeled data, eliminating the need for human annotation. In the following, we explore various self-supervised techniques to learn video representations, beginning with unimodal approaches based on pretext tasks.

2.3.1 Unimodal Self-Supervised Representation Learning

Pretext-Tasks Approaches

Early self-supervised methods for images followed the same paradigm as supervised approaches. Different types of corruption were applied to the images before feeding them to the network to recover the initial image. As

their name states, "pretext" tasks try to solve mock tasks that mimic children's puzzles. By solving those mock tasks, these models learn low-level local statistics about the extracted patch features in large datasets. Here is an example of the two different types of pretext tasks:

- **Data Imputation Tasks:** In these pretext tasks [Pathak 2016, Agrawal 2015, Zhang 2016], the original raw image is either partially occluded or a paired version from a different modality is utilized to predict the missing part of the image. Predicting color [Zhang 2016] has the advantage of using practically free training data: any color photo can serve as a training example by using the image's L (lightness value) channel as input and its ab (color value) channels as the supervisory signal. In [Agrawal 2015], the authors propose predicting camera transformation from image pairs as a pretext task, drawing inspiration from how humans perceive and learn egomotion. In [Pathak 2016], authors crop various parts of images, learning to predict and reconstruct the cropped windows.

- **Context Prediction Tasks:** Unlike the first tasks, spatial context tasks are designed as classification tasks where the model learns to classify the corruption done on it. For example, in [Doersch 2015], the authors consider predicting the relative position of sampled patches from an image as a pretext task. Two neighboring patches are sampled, and the relative position is predicted out of 8 possibilities.

Pretext tasks designed for 2D image processing can be adapted to 3D applications, enabling the training of 3D models using medical data such as MRIs. Authors of [Taleb 2020] have demonstrated the advantages of these pretraining objectives [Doersch 2015, Gidaris 2018] in brain MRI segmentation.

Similarly, pretext tasks exploiting the temporal dimension of videos have also been proposed [Wang 2020, Xu 2019, Misra 2016] The goal of these methods is to predict the corruption applied to the raw video by learning the transformation categorically through classification.

Applications in SDS Drawing from general computer vision advancements, researchers have adapted pretext tasks for the surgical domain to improve instrument localization and identification. In [Ross 2017], the authors pretrain a segmentation model on unlabelled data, using the colorization approach (see Fig. 2.4) from [Zhang 2016]. In [Yengera 2018], the authors propose using the remaining surgery duration as a pretext task for recognizing surgical phases. Tackling the temporal aspect of lengthy surgical video understanding, [Bodenstedt 2017] introduced a frame-sorting pretext task (see Fig. 2.5).

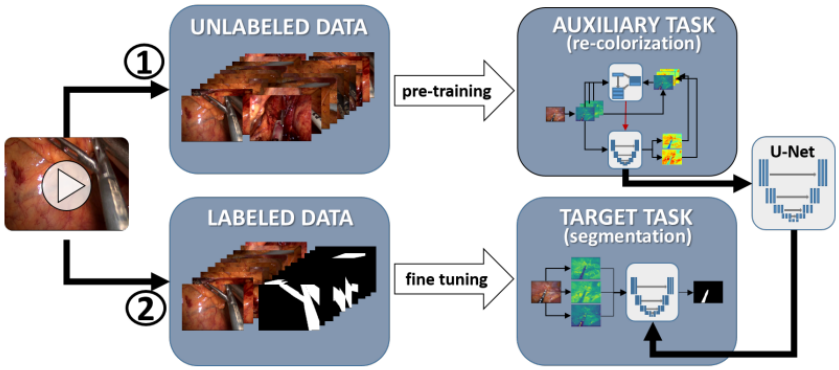

Figure 2.4: Overview of the approach presented in [Ross 2017]. This method involves a two-step process: pretraining through a colorization task on unlabelled data, then fine-tuning the model using available labelled data. Courtesy of [Ross 2017].

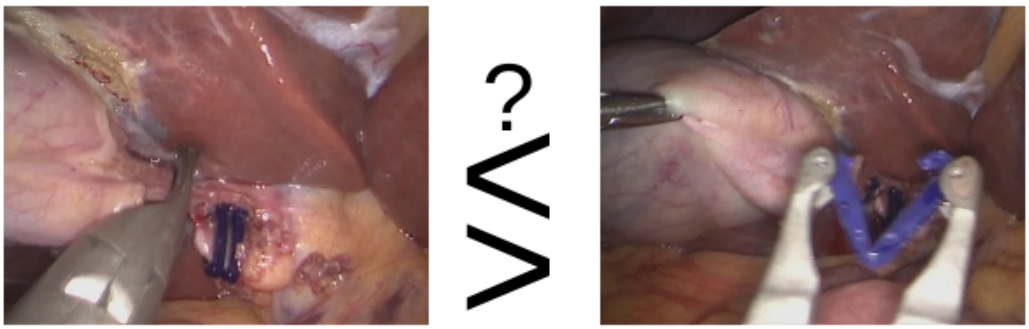

Figure 2.5: Overview of the approach presented in [Bodenstedt 2017]. This method involves a two-step process: pretraining is performed through a frame ordering task on unlabelled data and fine-tuning the model using available labelled data. Courtesy of [Bodenstedt 2017].

Contrastive Representation Learning

Contrastive learning is widely used in self-supervised learning and has achieved strong results in vision tasks [Chen 2020, He 2019, Caron 2020, Grill 2020]. It maps sample features onto a unit hypersphere, ensuring that positive sample pairs are close together while negative pairs are pushed apart. SimCLR [Chen 2020] used large mini-batch sizes for diverse negative samples. MoCo [He 2019] addressed the computational bottleneck with momentum encoders and a sample queue, while SwAV [Caron 2020] and BYOL [Grill 2020] opted to eliminate negative samples entirely.

Applications to Surgical Data Science In [Ramesh 2022], the authors conduct a comprehensive evaluation of four state-of-the-art SSL meth-

ods, MoCo v2 [He 2019], SimCLR [Chen 2020], SwAV [Caron 2020], and DINO [Caron 2021] within the context of surgical computer vision. Utilizing the Cholec80 dataset [Twinanda 2016], they assess the efficacy of these methods on two fundamental tasks: surgical phase recognition and tool presence detection. The study demonstrates that pretraining SSL models on endoscopic data, followed by fine-tuning with limited annotations, significantly enhances performance. Notably, they report improvements of up to 7.4% in phase recognition and 20% in tool detection over generic SSL applications, and up to 14% over existing semi-supervised baselines. This underscores the potential of in-domain SSL pretraining to reduce annotation requirements while maintaining high accuracy in surgical video analysis.

Generative Representation Learning

Drawing inspiration from the autoencoder-based pre-training methods used in Natural Language Processing, masked autoencoding techniques have also been proposed for image and video analysis. The authors of BERT (Bidirectional Encoder Representations from Transformers) [Devlin 2019] introduced a denoising autoencoder that operates on discrete representations. Given the textual information's compactness and semantic richness, the masking ratio is kept relatively low, at only 15 %.

Given the high redundancy of information in visual data, the masking ratio must be adjusted accordingly to prevent trivial reconstruction. Multiple masking strategies have been designed to make the masked regions more semantically meaningful. In [Sun 2022], a method has been proposed, focusing on the image's high-motion parts using different optical flow information derivatives. Other works [Wei 2022] preferred using the latent space for reconstruction, bootstrapping the power of strong SSL features.

Applications to Surgical Data Science In SurgMAE [Jamal 2023b], the authors introduce a modified MAE approach. They employ a targeted token sampling strategy that selects tokens from regions with high spatio-temporal activity, addressing limitations associated with conventional random masking techniques. The proposed method demonstrates improved performance on two surgical datasets [Hajj 2019, Sharghi 2020].

2.3.2 Multimodal Self-Supervised Representation Learning

Unlike unimodal representation learning, multimodal representation learning focuses on aligning the representations from different modalities in a shared embedding space. The primary objective of these alignment strategies is to enhance the performance of each modality individually while ensuring that the integration of both modalities improves performance on downstream tasks.

Multimodal pretraining also increases cross-modal retrieval and predictive coding abilities. By optimizing representations of each modality feature

in a shared latent space, multimodal representation learning allows for more robust representations resistant to noise and missing data [McKinzie 2023].

Contrastive Representation Learning

Similar to unimodal contrastive representation learning, methods are designed to optimize the similarity between paired instances using multiple modalities. One such method is CLIP [Radford 2021], which employs a dual-encoder approach, where each modality is processed by a separate encoder. CLIP creates a multi-modal embedding space through the joint training of an image and text encoder. CLIP aims to maximize the cosine similarity between paired instances while minimizing the cosine similarity between other instances, using the InfoNCE loss [van den Oord 2018].

Applications to SDS In their study, Jamal et al. [Jamal 2022] introduce a multimodal adaptation of the self-supervised clustering method proposed in [Caron 2020]. Instead of generating multiple augmented views from the same video frame or clip, the authors leverage synchronized depth images as complementary views, effectively incorporating depth information into the image encoder. They demonstrate the efficacy of their approach through data-efficiency experiments on two surgical workflow monitoring datasets, OR-AR [Sharghi 2020] and OR-Seg [Li 2020a], evaluating both activity recognition and semantic segmentation tasks using external camera viewpoints.

Cross-Modal Completion

Inspired by the method described in [Devlin 2019], which involves a transformer encoder-decoder architecture for predicting masked inputs in an autoregressive manner, the authors of [Lu 2019] propose an extension of this approach in a multimodal context. They mask semantic features from visual and textual branches and utilize cross-modal connections in the later layers to effectively fuse information from both modalities.

In CROC [Xie 2024], the authors explore a method for reconstructing masked visual features that have already been aligned. They use instruction tokens and unmasked visual features as inputs to an LLM to reconstruct the masked visual tokens. Instead of using the traditional unique mask token, they propose a dynamic token pool to replace the masked tokens. This approach provides better contextual cues and allows for a stronger masking ratio, which has yielded more effective results with larger ratios. MultiMAE [Bachmann 2022] adapts multimodal masking to diverse visual inputs, including depth, RGB images, and semantic segmentation maps. They maintain a low computational load by employing high masking ratios across different modality encoders. Their experiments demonstrated strong cross-modal coding capabilities and significant transfer performance improvements. Their method was later expanded in 4M [Mizrahi 2024] to

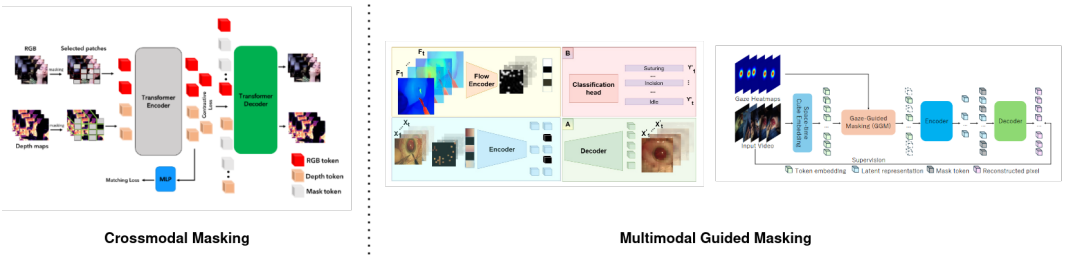

Figure 2.6: Architectures of two different multi-modal masking strategies [Jamal 2023b, Mostafa 2025]. **Right**: In M33D [Jamal 2023a] authors propose a joint RGB and Depth masking strategy. **Left**: authors of [Mostafa 2025] propose using a joint optical-flow RGB masked autoencoder.

include not just image-like modalities but also to utilize textual data and abstract modalities such as human pose and graph data.

Applications to SDS Concurrent with our prior research on OR-AR [Sharghi 2020], Jamal et al. proposed M^{3}3D [Jamal 2023a], a multimodal masked autoencoder that leverages synchronized intensity and depth maps. Integrating these modalities within the encoder facilitates the transfer of depth information into the intensity encoding, resulting in improved performance in OR-AR activity recognition tasks. In contrast, other studies [Mostafa 2025, Fujii 2024] adopted an alternative approach by adapting VideoMAE [Tong 2022] to utilize either optical-flow or gaze-based guidance for selective masking. Specifically, Mostafa et al. [Mostafa 2025] employed optical-flow guidance to mask image regions exhibiting substantial motion, compelling their model to concentrate on semantically meaningful regions within dynamic video clips. Similarly, in EgoSurgeryPhase [Fujii 2024], authors masked regions corresponding to areas of visual fixation, achieving comparable performance enhancements.

Multimodal Knowledge Distillation

Knowledge distillation was initially introduced by Hinton et al. [Hinton 2015] as a technique to transfer knowledge from a teacher model to a student model. This process involves training the student model to replicate the intermediate representations of the teacher model. Although it was initially designed for model compression, knowledge distillation has also been explored for cross-modal knowledge transfer, where the teacher and student models receive different types of input data [Gupta 2015, Aytar 2016].

Radevski et al. [Radevski 2023] focus on utilizing multimodal data exclusively for pretraining, demonstrating the effectiveness of their approach in unimodal downstream experiments. They also incorporate layout modality using object bounding box information during the pretraining phase. In their work on MM-CDFSL [Hatano 2024], the authors propose distill-

ing knowledge from pretrained multimodal masked encoders trained using RGB, optical flow, and hand pose information. They implement ensemble inference with three different teacher models, achieving state-of-the-art performance on action recognition benchmarks.

2.4 OR Workflow Monitoring

This section reviews literature on video representation learning with a focus on applications in surgical data science. The first section explores global appearance-based models for surgical video action recognition, highlighting the evolution from traditional machine learning to advanced deep learning methods. The second section examines structured semantic representations derived from abstract modalities, including object detection, human pose estimation, and superpixel segmentation, introduced previously (see Section 1.2.3). The third section discusses recent advances in self-supervised learning for surgical video analysis, addressing annotated data scarcity through unimodal and multimodal approaches.

Building upon these foundational methods, the following section synthesizes these representation learning techniques within the context of operating-room workflow monitoring. Specifically, it explores how the global, structured, and self-supervised methods described earlier have been adapted and integrated into systems capable of perceiving and analyzing complex OR environments, including human-pose estimation and automated detection of communication protocols derived solely from external camera views.

2.4.1 Human Pose Estimation and Activity Recognition in the OR

Occlusions, clutter, and variable lighting challenge HPE in the OR. Early works like [Kadkhodamohammadi 2014, Kadkhodamohammadi 2015] leveraged RGB-D inputs and 3D pictorial structures for robust clinician tracking. The release of datasets such as TUM-OR [Belagiannis 2016] and MVOR [Srivastav 2018] enabled learning-based, multi-view pose estimation. Unsupervised domain adaptation techniques, notably AdaptOR [Srivastav 2021], reduced reliance on labeled data and enabled privacy-preserving models. In parallel to these works focusing on HPE in the OR, in [Twinanda 2015], authors propose an action classification algorithm using STIP [Laptev 2003] and HOG features to classify OR video clips into the proper action class. Surgical activity recognition was later studied in [Sharghi 2020], by using a two-stage approach with 3D convolutional networks to extract features on the large OR-AR dataset [Sharghi 2020], and then training GRU models to model long-range temporal dependencies. Recent work extends this line through multi-view fusion [Schmidt 2021] and self-supervised training regimes [Jamal 2023b, Jamal 2022].

2.4.2 Human-Object Interactions and Scene Understanding

Accurately modeling staff–instrument interactions in the OR requires rich semantic understanding of the scene. Early work by Ladikos et al. [Ladikos 2008] addressed this problem by reconstructing a 3D representation of all objects to predict potential collisions. More recent multi-camera systems extend this idea, enabling simultaneous tool tracking and surgeon-hand assignment [Basiev 2021]. Researchers have since broadened the scope from instruments to full-scene semantic segmentation, labeling every clinically relevant entity—personnel, instruments, operating tables, ceiling lights, monitors, and more—to build holistic context maps. For robotic ORs, Li et al. [Li 2020a] proposed a multi-view 3-D segmentation framework, while SegmentOR [Bastian 2023b] accelerates annotation through temporal label propagation. Scene-graph approaches such as 4D-OR and LABRADOR [Özsoy 2022, Özsoy 2023] further structure these semantics, encoding object relationships and procedural roles for higher-level reasoning. Finally, gaze-estimation techniques [Gershov 2022] complement physical tracking by capturing attention patterns. This capability can be integrated into scene graphs to refine our understanding of fine-grained interactions and communication protocols within the OR.

2.4.3 Skill Assessment and Protocol Recognition

Recent work focuses on evaluating technical skills and verifying safety protocol compliance. Depth-based methods [Zuckerman 2024] enable robust, anonymous skill metrics extraction by detecting surgeon hands and surgical tools. Multi-modal OR data supports automatic phase and gesture classification [Basiev 2021]. Protocol adherence detection, such as recognizing ”Time-Out” and ”StOP?” events [Chen 2025], leverages multi-camera video and team activity modeling. On the protocol side, advanced video analytics can verify if safety protocols like Time-Outs are conducted, helping to improve compliance and safety. These developments close the loop in OR workflow monitoring: We can observe what and who is in the OR and evaluate how well the team is performing and whether they are following the critical steps safeguarding patient outcomes. The continuing challenge lies in generalizing these models to diverse surgical procedures and hospital settings and integrating them smoothly into the OR without disrupting routines. Nonetheless, the progress reviewed in this section demonstrates a clear path toward smarter ORs where AI-driven systems support surgeons and staff by providing timely insights into workflow dynamics, skill levels, and protocol adherence, ultimately aiming for higher efficiency and improved patient safety.

2.5 Thesis Positioning

The primary objective of this thesis is to explore how self-supervised learning methods can benefit from structured video representations based on *abstract* modalities in surgical data science. In contrast to prior SSL studies that operate solely on raw pixel streams, we develop multimodal pretraining objectives over semantically rich abstractions, object layouts, human poses, and over-segmented visual regions to improve downstream understanding of surgical workflows.

This research is motivated by three major challenges in SDS:

- the high cost of data annotation
- the variability in camera angles and frequent occlusions in operating rooms
- the importance of capturing human-object and human-human interactions for fine-grained workflow understanding

Structured representations offer a natural way to address these challenges by directly embedding prior knowledge about scene semantics into the learning process.

Our contributions begin with a detailed study of two multimodal OR datasets [Li 2020a, Sharghi 2020] focused on semantic segmentation and activity recognition, respectively. We propose a novel pretext task combining unsupervised superpixel segmentation with depth-based privacy-preserving representations. This task is designed to operate on local visual features extracted from depth maps.

Furthermore, we explore the integration of structured modalities within contrastive, generative, and multimodal SSL frameworks. In contrast to previous work [Jamal 2022, Jamal 2023b, Jamal 2023a], our models are trained to align and reconstruct information from abstract representations rather than raw frames alone. This structured supervision improves label efficiency and provides more interpretable visual understanding.

This is the first comprehensive study systematically combining structured video abstraction with modern self-supervised learning paradigms for SDS. The outcomes of this thesis pave the way for more data-efficient and semantically grounded solutions in surgical workflow analysis using external cameras.

Part I

Contributions

3

Analysis of Two Operating Room Workflow Monitoring Datasets

Contents

3.1 Introduction

The availability of large-scale annotated video datasets has significantly accelerated advancements in automatic video analysis. Popular general-purpose datasets capturing daily-life activities, such as Something-Something [Goyal 2017], Kinetics [Carreira 2017], and UCF101 [Soomro 2012], have contributed substantially to progress in computer vision but often exhibit limited variability due to controlled recording environments.

In medical domains, particularly surgical data science, privacy constraints have restricted the creation of similarly extensive and diverse datasets. Recent efforts in surgical vision have produced specialized datasets for various tasks including surgical phase recognition (e.g., Cholec80 [Twinanda 2016], CATARACTS [Hajj 2019], MultiBypass140 [Lavanchy 2023a]) and semantic segmentation (e.g., EndoVis [Allan 2019, Allan 2020], CholecSeg8K [Hong 2020], Endoscapes [Murali 2022]). However, these datasets predominantly represent internal surgical views and do not adequately capture the operating room workflow. This is critical for understanding human activities and interactions within the OR.

Initial work on room-level understanding by Twinanda et al. [Twinanda 2015] introduced a private, multi-view RGB-D dataset comprising fifteen atomic actions in vertebroplasty. Subsequent annotations aggregated these primitives into higher-order activities, yet the data remain unavailable to the community. Public initiatives have primarily concentrated on human-pose estimation: MVOR and TUM-OR provide multi-view RGB-D sequences with clinician keypoints and bounding boxes [Srivastav 2018, Belagiannis 2016]. While indispensable for pose research, they include limited or no activity labels, marking OR-level action recognition as underrepresented.

To bridge this gap, this thesis focuses on analyzing the OR-AR dataset [Sharghi 2020], aimed explicitly at detecting human activities in the OR environment across robotic-assisted and traditional surgeries. Additionally, a subset of OR-AR is extended with annotations for detecting OR devices and clinicians. We also present OR-Seg [Li 2020a], a multiview semantic segmentation dataset specifically focusing on robotic-assisted surgical instruments, leveraging depth map recordings captured in OR environments.

3.2 OR-AR Dataset

To automatically recognize the different activities related to OR workflow monitoring, one must first describe them and their critical aspects to annotate them properly without confusion. Here, we will define the 10 activities annotated in the OR-AR dataset [Sharghi 2020, Jamal 2023b].

3.2.1 Activity Definitions

- **Sterile Preparation** refers to all preoperative measures taken to establish and maintain a sterile surgical environment before the first incision [Kanji 2021]. This phase includes preparing the operating room and instruments: for example, draping the robotic arms with sterile covers, organizing and opening sterile instrument trays, and ensuring all needed equipment is sterile and in place.

- **Patient Roll-In/Roll-Out** Roll-In is the process of bringing the patient into the operating room and transferring them onto the operating table at the start of the procedure. This typically coincides with the initiation of anesthesia [Kanji 2021]. Conversely, patient roll-out refers to transferring the patient off the operating table and out of the OR at the end of the procedure [Zamudio 2023]. This final transfer occurs after surgery, complete wound closure, and the patient has been awakened from anesthesia.

- **Patient Preparation** involves properly positioning and securing the patient. The surgical site on the patient's body is then sterilized (skin antiseptic prep) and draped with sterile sheets, integrating the patient into the sterile field [Kanji 2024].

- **Robot Roll-Up/Roll-Out** denotes positioning the robotic surgical system at the patient's bedside in preparation for docking. Once the patient is prepped, the OR staff rearrange equipment and clear a pathway for the robot to approach the operating table. Robot roll-out is the step of pulling the robotic cart back and away from the operating table. In other words, the robot is wheeled out of the immediate operative field to allow unobstructed access to the patient for the final steps [Kanji 2021].

- **Robot Docking/Undocking:** Robot docking is the process of attaching the robotic system to the patient's anatomy via the surgical trocars (ports). After the robot has been rolled into place, each robotic arm is "docked" by connecting it to a port inserted into the patient. The camera arm is docked first, followed by the robotic arms. Robot undocking is performed at the end of the robotic portion of surgery, essentially detaching the robot from the patient. Robot Undocking requires removing all robotic instruments and disengaging each arm from the trocars in the patient in a controlled sequence.

- **Robotic Surgery** is the main intraoperative period the surgeon operates from the robotic console. After docking, the surgeon (and any console-assisting surgeons or trainees) physically step away from the

Figure 3.1: Visualization of the ten workflow monitoring activities annotated on OR-AR [Sharghi 2020], with their corresponding label occurrences. Courtesy of [Sharghi 2020]

patient's bedside to the console to begin the robotic procedure [Zamudio 2023].

- **Patient Close:** refers to all the activities in concluding the operation after the robotic work is done, particularly closing and securing the surgical incisions. Once the robot is undocked and moved aside, the surgical team closes any internal fascial layers and then sutures or staples the skin incisions used for the ports or assist openings.

In [He 2022b, Jamal 2023b], authors discard the *sterile preparation* activity in their activity recognition benchmarks. We follow the same evaluation protocol in [Hamoud 2023, Hamoud 2025].

3.2.2 Acquisition of Videos

Two imaging carts were placed in each room, for a total of four carts, each equipped with two ToF cameras. The baseline between the cameras on each cart is 70 centimeters, and their orientation is fixed. This results in a slightly different view in videos captured by the cameras from the same cart. However, different carts in the same room are set in strategic positions, such that if a cart's view is blocked due to clutter in the scene, the other cart can successfully capture the activities. The OR-AR dataset contains 400 full-length videos extracted from 103 surgical procedures. Videos were taken using time-of-flight sensors and placed around the OR during robotic surgery. The data was collected from 27 surgeons and included 30 types of surgeries across two ORs over two years. The ToF cameras acquire intensity images

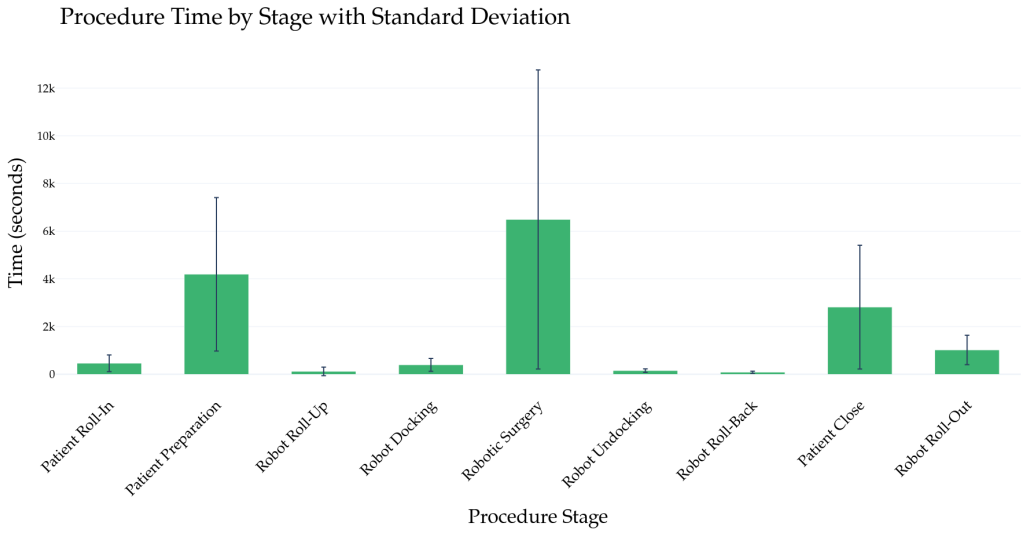

Figure 3.2: A breakdown of the different activity durations showing average duration and associated standard deviation.

based on the reflection of infrared rays and depth maps of the recorded scene.

3.2.3 OR-Det Dataset

The OR-AR dataset was later extended with bounding box annotations for persons and objects. Five OR-specific objects and clinicians are annotated on around 19K frames across 20 full-length videos. Objects that were annotated with bounding boxes were: **surgical gurney**, **sterile/non-sterile table**, **patient side-cart (PSC)**, **OR table**, **vision side-cart (VSC)**.

3.2.4 Evaluation Metrics

Action Recognition and Activity Recognition We distinguish between two related but distinct tasks: action recognition and activity recognition. Action recognition, commonly defined within the computer vision field, refers to classifying short video clips into action categories. Typically, these clips illustrate a single action. In contrast, activity recognition involves segmenting a more extended video sequence into relevant temporal segments, each representing a specific activity containing several actions. Within the SDS community, this task is also known as phase recognition.

- **Action recognition (clip classification):** Short clips are sampled so that each lies entirely within a single activity segment (e.g., a 16-second excerpt taken from the *Robot Docking* portion of the procedure). Because every clip corresponds to exactly one label from the shared vocabulary, the task is a standard multi-class classification problem. Performance is reported with top-k accuracy, where a prediction is

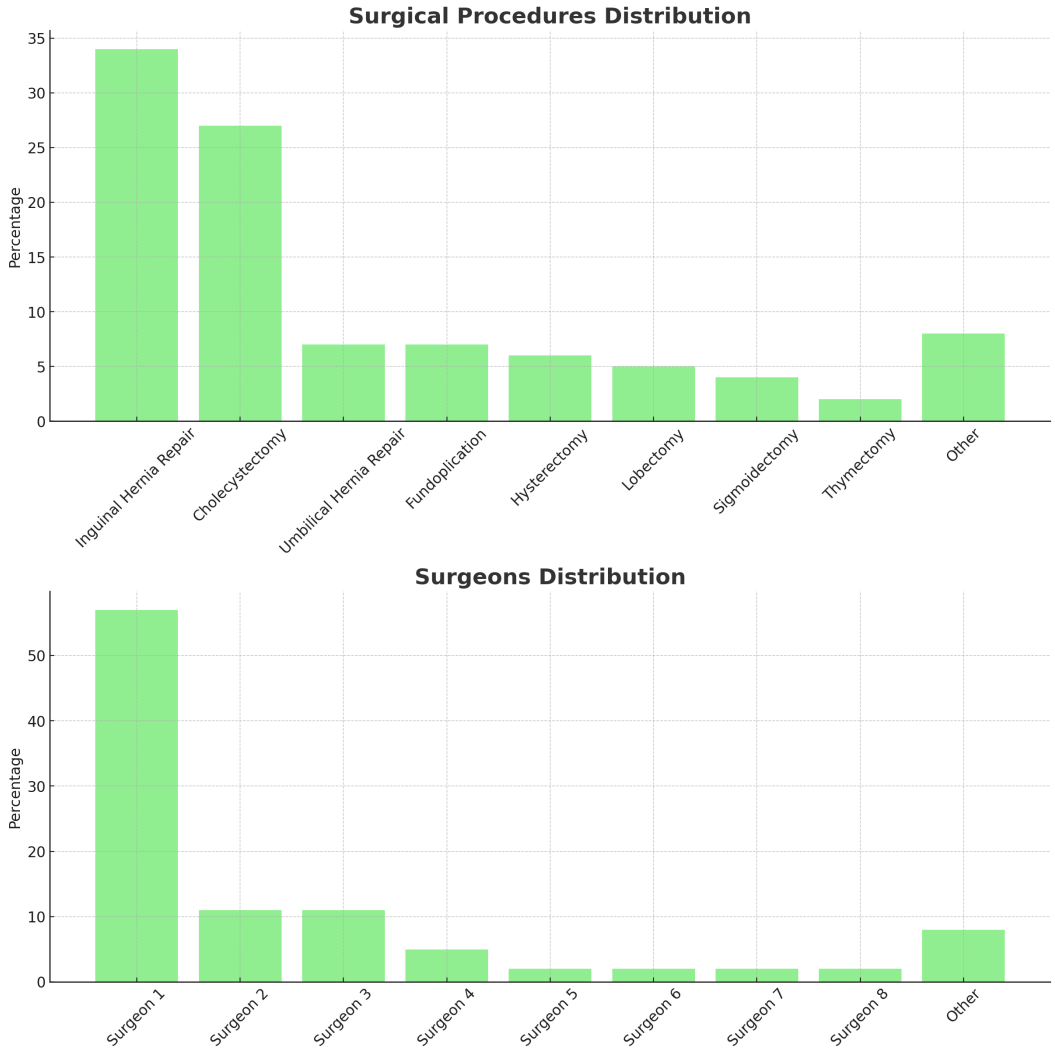

Figure 3.3: **Above:** A breakdown of the type of procedure distribution in the OR-AR dataset. **Below:** A breakdown of the surgeon distribution in the OR-AR dataset.

Figure 3.4: Components of the da Vinci Surgical System: Patient Side Cart, Vision Side Cart, and Surgeon Console. (Image courtesy of [Avgousti 2020])

deemed correct if the ground-truth label appears among the model's k highest-scoring classes.

- **Activity recognition (long-video segmentation):** A continuous OR recording is processed with a sliding window: each window is first classified by the action-recognition model, after which temporal modeling (e.g., GRU [Chung 2014] decoding) produces contiguous time segments. The resulting timeline is expressed with the *same* label set, but now each label spans a variable-length interval. Following prior work, segmentation quality is summarized by the mean Average Precision (mAP) over all activity classes [Sharghi 2020, He 2022b].

Object and Person Detection We use the Average Precision (AP) $AP_{0.5:0.95}$ metric from COCO for the evaluation, which is the average over multiple IoU (the minimum IoU to consider a positive match) from 0.5 to 0.95 with a step of 0.05. We also provide $AP_{0.5}$ and $AP_{0.75}$ as extra metrics to evaluate the different baselines' performance at different precision levels in Table 3.2 and Table 3.1.

3.2.5 Results

Benchmark of SOTA Object Detectors The proposed frameworks in Chapter 5 rely on detected objects and clinicians in the form of bounding boxes on the integrity of the OR-AR dataset. As such, we preferably run SOTA object detectors [He 2017, Carion 2020a, Zhu 2020] to extract layout information for our object-centric frameworks. The statistics of our detected objects and the performance of our surgical and clinician detectors are provided in Tables 3.2 and 3.1.

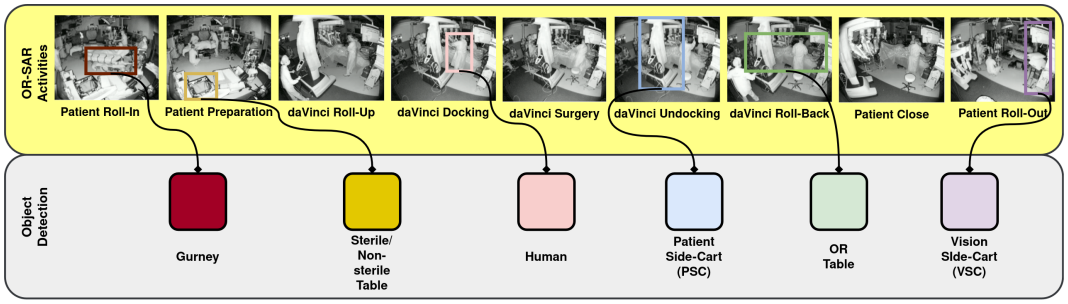

Figure 3.5: Visualization of the nine annotated activities in the restricted OR-AR dataset, with the annotated objects from the OR-Det dataset [Hamoud 2023].

Backbone	**Method**	**Dataset**	*AP*	*AP@0.5*	*AP@0.75*
ResNet50	Mask-RCNN	COCO	76.8	94.0	86.5
ResNet101	Mask-RCNN	COCO	78.3	95.3	87.7
ResNet50	DefDETR	COCO	76.9	**96.3**	87.3
ConvNeXt-B	C-Mask-RCNN	IN1K	**79.6**	**96.3**	**89.5**

Table 3.1: Performance of different object detection methods for clinician detection. Performance is given in (%). Results are given after fine-tuning on OR-Det.

Backbone	**Method**	**Dataset**	AP	AP@0.5	AP@0.75
ResNet50	Mask-RCNN	COCO	59.3	86.0	61.6
ResNet101	Mask-RCNN	COCO	60.0	87.6	62.4
ResNet50	DefDETR	COCO	59.7	87.2	62.1
ConvNeXt-B	C-Mask-RCNN	IN1K	**61.8**	**89.3**	**64.1**

Table 3.2: Performance of different object detection methods for surgical device detection. Performance is given in (%). Results are given after fine-tuning on OR-Det.

Benchmark of SOTA Activity Recognition Models When introducing the OR-AR dataset, authors also proposed a comprehensive benchmark of different clip classification backbone models associated with temporal models. We provide their results in Table 3.3.

Table 3.3: Performance comparison (%) of different backbone and temporal model combinations.

Backbone	**Transformer**	**Bi-GRU**	**Uni-GRU**	**TCN**
I3D	79.30±0.06	94.04±0.66	90.95±0.74	91.33±0.23
SlowFast	79.42±1.71	94.33±0.19	90.70±0.4	89.79±1.08
TimeSformer	76.23±0.33	93.20±0.04	88.89±0.66	89.59±0.07
Swin	**82.50±2.35**	**95.13±0.35**	**92.02±0.69**	**91.54±0.03**

3.3 OR-Seg Dataset

Precise localization of robotic surgical devices is crucial for developing context-aware systems. For example, the pixelwise 3D localization of the Patient-Side Cart (the part of the surgical robot with robotic arms) could pave the way for smarter surgical robots to dock themselves automatically. As such, developing semantic segmentation tools in the robotic OR would be a necessary step towards semi-autonomous surgical robots.

3.3.1 System Components of the da Vinci Surgical System

The OR-Seg [Li 2020a] provides detailed annotations of the da Vinci Surgical System's components, focusing on the Patient-Side Cart (PSC) and the Vision-Side Cart (VSC). These annotations are instrumental in developing advanced perception systems for enhanced situational awareness in robotic-assisted surgical environments.

Patient Side Cart The PSC is the operative component of the da Vinci Surgical System, positioned adjacent to the patient during procedures. It comprises multiple robotic arms that manipulate surgical instruments and an endoscopic camera. Each arm has setup joints and instrument arms designed to establish a remote center of motion, allowing precise instrument maneuvering while minimizing force on the patient's body wall. In the dataset, PSC annotations include the spatial configuration of its robotic arms.

Vision Side Cart The VSC serves as the visual interface of the da Vinci Surgical System, housing the imaging and processing equipment necessary for providing the surgeon with a high-definition, 3D view of the surgical area. It includes components such as the illuminator, endoscopes, stereo camera head, camera control units (CCUs), and a touchscreen monitor. Annotations of the VSC in the dataset encompass its placement within the op-

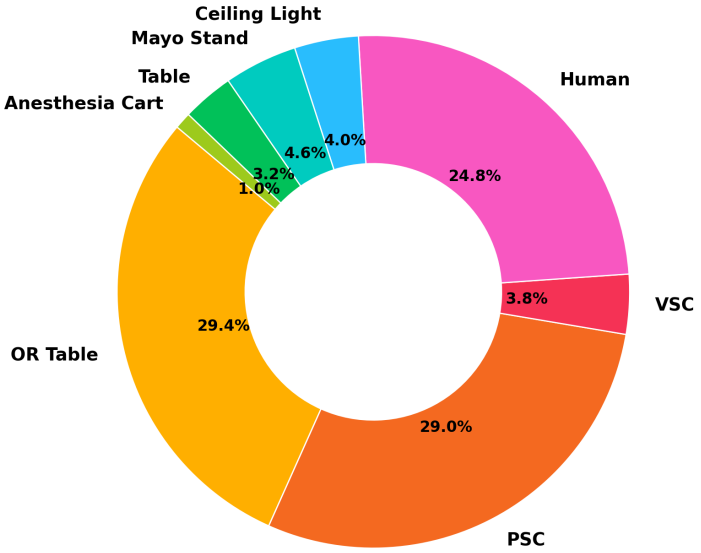

Figure 3.6: A breakdown of the distribution of pixelwise annotations for each of the eight annotated objects, discarding background.

erating room, the endoscopic camera's orientation and field of view, and the interconnections between its various components.

3.3.2 Acquisition of Videos

The data was collected in a clinical development lab, where different robot-assisted laparoscopic procedures were simulated, and the ToF cameras took videos. The salient frames, i.e., frames with significant enough differences, are then extracted from the videos. The dataset has two portions: single-view and multi-view. The single-view dataset consists of 7980 images. The data is captured by attaching the ToF cameras on the PSC directly, as shown in Fig. 3.7. The example images are shown in Fig. 3.7. The color code and pixel frequency for the eight classes in the dataset (excluding background) are shown in Fig. 3.6.

3.3.3 Evaluation Metrics

Semantic Segmentation We use the standard mean Intersection over Union (mIoU) metric for our semantic segmentation task. In [Li 2020b], authors also provide frequency weighted Intersection over Union (fwIoU), which we will also provide in Table 3.4.

3.3.4 Results

The authors also proposed a comprehensive benchmark of different semantic segmentation models when introducing the OR-Seg dataset. We provide their results in Table 3.4.

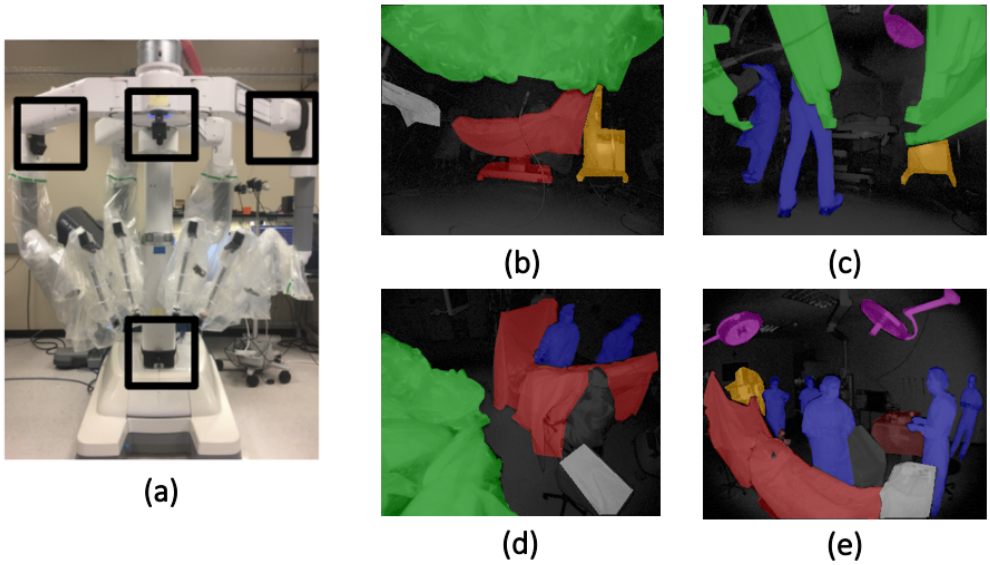

Figure 3.7: (a) PSC robot with ToF cameras attached (in black rectangles) (b-e) OP, USM1, USM4, BASE Camera viewpoints from the multi-view dataset. Courtesy of [Li 2020a].

	mIoU	fwIoU
DeepLab V3+ [Chen 2017]	0.670 ± 0.036	0.857 ± 0.014
CRF [Krähenbühl 2011]	0.671 ± 0.036	0.859 ± 0.014
MVPM [Li 2020a]	0.685 ± 0.028	0.860 ± 0.011

Table 3.4: Comparison of mIoU and fwIoU for different models. Courtesy of [Li 2020a]

3.4 Conclusion

In summary, this chapter has characterised the two datasets that underpin all subsequent experiments in this thesis, OR-AR [Sharghi 2020] for multiview workflow monitoring and OR-Seg [Li 2020a] for pixel-accurate segmentation of robotic devices. By comparing them with existing resources (Table 3.5), we have shown that they uniquely combine depth imagery, room-level activity labels, and detailed object annotations, filling critical gaps in surgical-vision benchmarks. The algorithms developed in the following chapters will leverage these datasets to advance state-of-the-art performance in activity recognition and semantic segmentation, ultimately contributing to safer and more autonomous surgical workflows.

Table 3.5: Comparison of Operating Room Datasets for Segmentation and Activity Recognition

Dataset (Year)	Modalities	Cams	Size	Annotation	Real Proc.
VerCArm24 [Twinanda 2015]	RGB-D (3× Asus Xtion), synchronized views	3	1734 multiview video clips	Action/Phase labels	✓
sensORs [Bastian 2023a]	RGB-D (2× Kinect), synchronized views	2	16 videos 24 hours	Action/Phase labels	✗
TUM-OR [Belagiannis 2016]	RGB (5×), synchronized views	5	7000 multi-view frames	2D/3D human pose estimation	✗
MVOR [Srivastav 2018]	RGB-D (3× Kinect), synchronized views	3	732 multi-view frames	2D/3D human pose estimation	✓
4D-OR [Özsoy 2022]	RGB-D, point clouds from simulated OR	6	6,734 timepoints	3D boxes, scene graphs, role labels	✗
MM-OR [Özsoy 2025]	RGB-D video, audio, transcripts, robot logs	8	92,983 total; 25,277 labeled	Panoptic segmentation, scene graphs, temporal labels	✓
OR-AR [Sharghi 2020]	ToF depth, IR intensity	4	≈100 hours from 103 procedures	Temporal phase/activity labels	✓
OR-Seg [Li 2020a]	ToF depth, IR intensity	4	7,980 single-view; 100 multi-view	Pixel-wise 2D semantic segmentation	✓

4

Depth-based OR Workflow Monitoring with Superpixel Self-Supervision

Contents

Figure 4.1: Visualization of a surgical scene with superpixel clusters reprojected and colorized on the point cloud.

This chapter introduces a novel self-supervised learning framework designed to achieve efficient, privacy-preserving context awareness in robot-assisted surgical environments. Our approach leverages spatial depth information captured by Time-of-Flight cameras to minimize dependency on manual annotations. Extensive benchmarking against established self-supervised methods, including RotNet [Gidaris 2018] and CPC v2 [van den Oord 2018], demonstrates superior data efficiency and performance across two publicly available datasets. This chapter is adapted from the following publication:

HAMOUD, I., KARARGYRIS, A., SHARGHI, A., MOHARERI, O., PADOY, N. (2022). SELF-SUPERVISED LEARNING VIA CLUSTER DISTANCE PREDICTION FOR OPERATING ROOM CONTEXT AWARENESS. IPCAI-INTERNATIONAL JOURNAL OF COMPUTER ASSISTED RADIOLOGY AND SURGERY, 17(8), 1469-1476.

4.1 Introduction

While robot-assisted surgery can enhance surgical outcomes, it also introduces additional complexities into operating room (OR) workflows, increasing the potential for procedural errors [Sheetz KH 2020, Catchpole 2015]. Recent technological advances enable the integration of sensor-based data collection within ORs, facilitating context-aware frameworks that have the

potential to significantly improve workflow understanding and surgical team coordination [Dias 2020].

Current computer vision and deep learning approaches have successfully enabled detailed three-dimensional (3D) scene understanding in ORs [Srivastav 2018, Sharghi 2020, Li 2020b], yet heavily rely on manual expert annotations. These annotations present substantial practical limitations due to their high time and resource demands.

Addressing this limitation, we propose a self-supervised method based on predicting relative Euclidean distances between superpixel regions in depth images. Our approach is implemented through a two-stage encoder-decoder architecture. In the first stage, we introduce an innovative pretext task that learns viewpoint-invariant spatial relationships inherent to depth data, crucial for reliable surgical context awareness. This viewpoint invariance arises naturally from the consistent spatial distances between anatomical and operational structures, irrespective of camera positioning.

In the second stage, we fine-tune the pretrained encoder for two surgical perception tasks: activity classification and semantic segmentation, demonstrating substantial data efficiency by reducing labeled data requirements progressively. To validate our approach comprehensively, we benchmark against established self-supervised techniques RotNet [Gidaris 2018] and CPC v2 [van den Oord 2018], adapting these methodologies specifically for depth data from ToF cameras.

The main technical contributions of this chapter are:

- Introduction of a novel viewpoint-invariant, self-supervised pretext task specifically tailored for depth data obtained from ToF cameras, addressing both data annotation challenges and privacy concerns
- Comprehensive experimental validation and benchmarking on two public surgical datasets [Sharghi 2020, Li 2020b], confirming our method's superior effectiveness and generalizability compared to current self-supervised learning approaches.

4.2 Methodology

4.2.1 Proposed Pretext Task

The robotic operating room is a highly streamlined platform in which the personnel and objects are expected to follow a certain protocol and be at a specific place at a specific time. Relative positions of objects in the room can provide powerful information to integrate into surgical workflow analysis. In contrast to other pretext tasks, such as the one proposed in [Doersch 2015], where only the 2D relative position of patches is used, our pretext task aims at taking advantage of the depth information by predicting the relative distance of objects in 3D without any annotations. To this

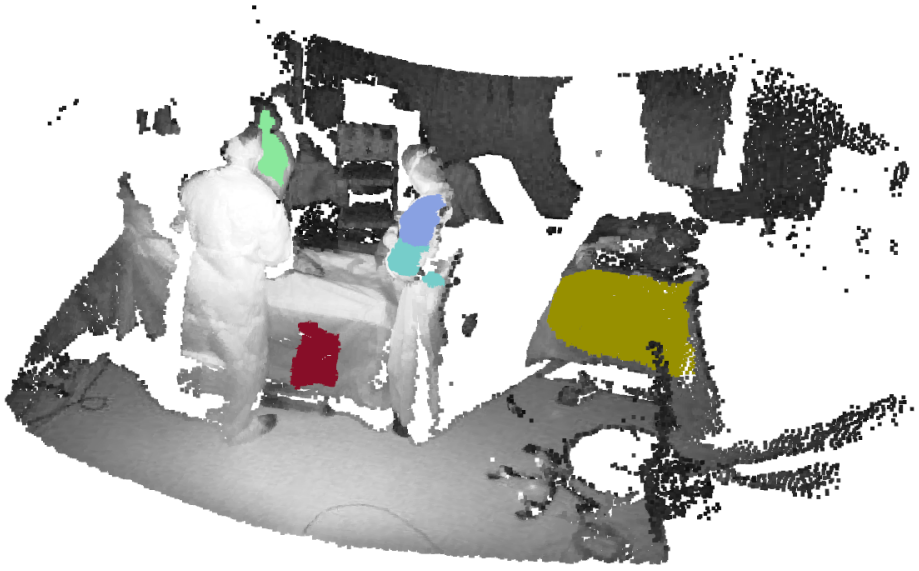

Figure 4.2: Selected Superpixels Displayed After Filtering.

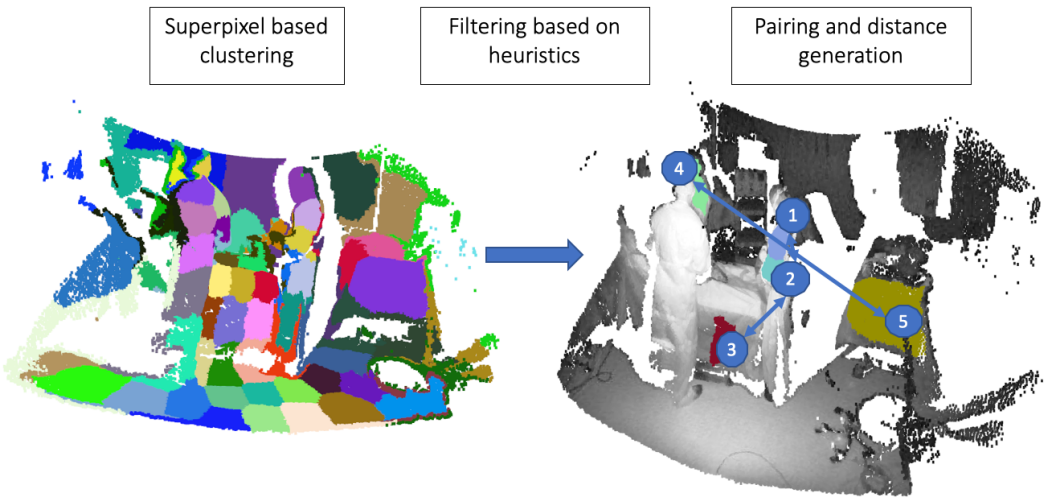

Figure 4.3: Pretext task annotation generation process using SLIC [Achanta 2012] superpixel segmentation.

end, we propose a new sampling method to extract homogeneous clusters from our depth map (see Fig. 4.1) so each cluster belongs to one specific object. Depth maps from different views are analyzed independently in our work. Inspired by *Ouyang et al.*, we propose a superpixel-based approach to compute our clusters. Superpixels tend to be small-scale, dense image regions that offer a nice and smooth unsupervised segmentation [Felzenszwalb 2004, Achanta 2012]. In this work, we employed the Simple Linear Iterative Clustering (SLIC) method [Achanta 2012] because it is faster and more memory efficient than other existing methods. In addition to these quantifiable benefits, SLIC is easy to use and offers flexibility in terms of compactness and the number of superpixels it generates.

4.2.2 Self-supervised labeling strategy

Superpixel Generation To extract homogeneous regions from our depth map, SLIC [Achanta 2012] is used to define regions of an overall identical depth and thus likely belonging to the same object (see Fig. 4.6). In our experiments, we used the scikit-image implementation of SLIC. SLIC algorithm clusters pixels in the five-dimensional space defined by the CIELAB color components (l, a, b) and the pixel coordinates (x, y). We convert our depth map information by repeating the L value 3 times for the RGB channels. The distance measure D used in SLIC combines color similarity and spatial proximity, and is defined as:

$$D = \sqrt{d_{lab}^2 + \left(\frac{m}{S} \cdot d_{xy}\right)^2} \tag{4.1}$$

where:

- $dlab$ is the Euclidean distance between the color vectors of the cluster center k and pixel i in the CIELAB color space:

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \tag{4.2}$$

- dxy is the Euclidean distance between the spatial coordinates of the cluster center k and pixel i:

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{4.3}$$

- m is the compactness factor that limits adjacency degree and controls the overall irregularity of superpixel shape.

- S is the grid interval, approximately equal to $\sqrt{HW/K}$, where H and W are the image dimensions, and K is the desired number of superpixels.

Algorithm 1: SLIC with Gaussian Smoothing

Input: Image I of size $H \times W$, number of superpixels K, compactness m, Gaussian smoothing parameter σ
Output: Label map L with superpixel assignments

Apply Gaussian smoothing to image I: $I_{\text{smooth}} \leftarrow G_\sigma * I$;
Compute grid interval: $S \leftarrow \sqrt{HW/K}$;
Initialize cluster centers C_k on a regular grid with spacing S;
Move each C_k to the lowest gradient position in a 3×3 neighborhood;
Initialize $L(i) \leftarrow -1$, $D(i) \leftarrow \infty$;
repeat
  foreach *cluster center* C_k **do**
    foreach *pixel* i *in* $2S \times 2S$ *region around* C_k **do**
      Compute color distance:

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2}$$

      Compute spatial distance:

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}$$

      Compute combined distance:

$$D_s(i) = \sqrt{d_{lab}^2 + \left(\frac{m}{S} \cdot d_{xy}\right)^2}$$

      if $D_s(i) < D(i)$ **then**
        $D(i) \leftarrow D_s(i)$;
        $L(i) \leftarrow k$;
  foreach *cluster* k **do**
    Update C_k as the mean of all assigned pixels;
until *convergence or max iterations*;
Enforce connectivity by relabeling disconnected components;

The approximate number of segments to be generated by SLIC was chosen as 500, based on experiments and qualitative training data analysis. We also choose $\sigma = 3$ as width of the Gaussian smoothing kernel for preprocessing the image to consider the noise in the depth maps. The compactness is chosen as $m = 3$ to balance the importance between spatial proximity in the image coordinates and depth proximity in the image intensity.

Filtering noisy clusters & Generating Pseudo-labels Once the superpixels are generated, we filter them using heuristics based on the convexity of

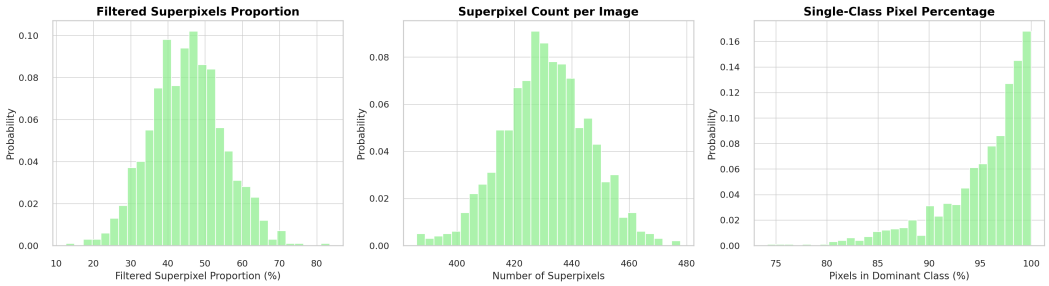

Figure 4.4: Statistics on (a) proportion of filtered superpixels per image, (b) number of superpixels per image, (c) percentage of pixels belonging to a single class in each superpixel on OR-Seg [Li 2020a]

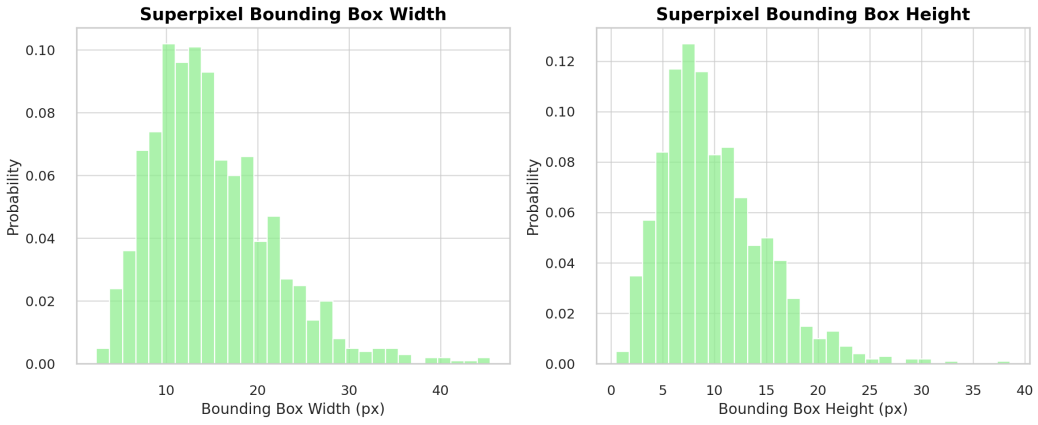

Figure 4.5: Distribution of width and height of superpixel clusters on OR-Seg [Li 2020a].

the superpixel region (solidity over 0.75), the disparity of the depth inside the cluster (std under 0.2m), and the number of missing values in the cluster (less than 5%) to prevent having regions with too much noise. These heuristics were chosen based on preliminary studies on our training data. We are only interested in compact regions with very low deviation in the three spatial directions. These superpixels can then be mapped to a set of points in the corresponding point cloud. The available camera intrinsics and the depth information are used to compute the corresponding coordinates for each point in our clusters. In the end, we obtain for each image I a set of point clusters $\{SP_1, ..., SP_N\}$ and define the distance between two clusters as the euclidean distance between the two centroids of the two point clouds (see Fig. 4.6):

$$SP_1 = \{x_1^1, ..., x_{N_1}^1\} \quad SP_2 = \{x_1^2, ..., x_{N_2}^2\} \quad x_i^j \in \mathbb{R}^3 \tag{4.4}$$

$$C_1 = \frac{\sum_{i=1}^{N_1} x_i^1}{N_1} \quad C_2 = \frac{\sum_{i=1}^{N_2} x_i^2}{N_2} \tag{4.5}$$

$$SP_{dist} = \|C_1 - C_2\|_2 \quad . \tag{4.6}$$

The euclidean distance between the superpixel clusters is regressed on the distance between the corresponding learned representations in the feature space, as expressed below:

$$h_{SP_1} = f_{extract}(D_{SP_1}) \quad h_{SP_2} = f_{extract}(D_{SP_2}) \tag{4.7}$$

$$l_2 = \|h_{SP_1} - h_{SP_2}\|_2 \tag{4.8}$$

$$L_{pretext} = \|l_2 - SP_{dist}\|_1 \quad . \tag{4.9}$$

where h_{SP_i} represents the feature vector extracted from our network $f_{extract}$ by giving the corresponding depth map patch D_{SP_i} as the input. $L_{pretext}$ is the loss regressed by our network illustrated in Fig. 4.6 and described in the next section.

4.2.3 Encoder-Decoder architecture

Feature Extraction: ResNet-50 [He 2015] has been successfully employed in many works for both semantic segmentation and activity detection. In this work, we also utilize the same architecture as our backbone visual feature extraction model. This model maps $224 \times 224 \times 1$ depth maps to a feature space of size $9 \times 9 \times 2048$. It is trained on frames extracted from the videos without any temporal context.

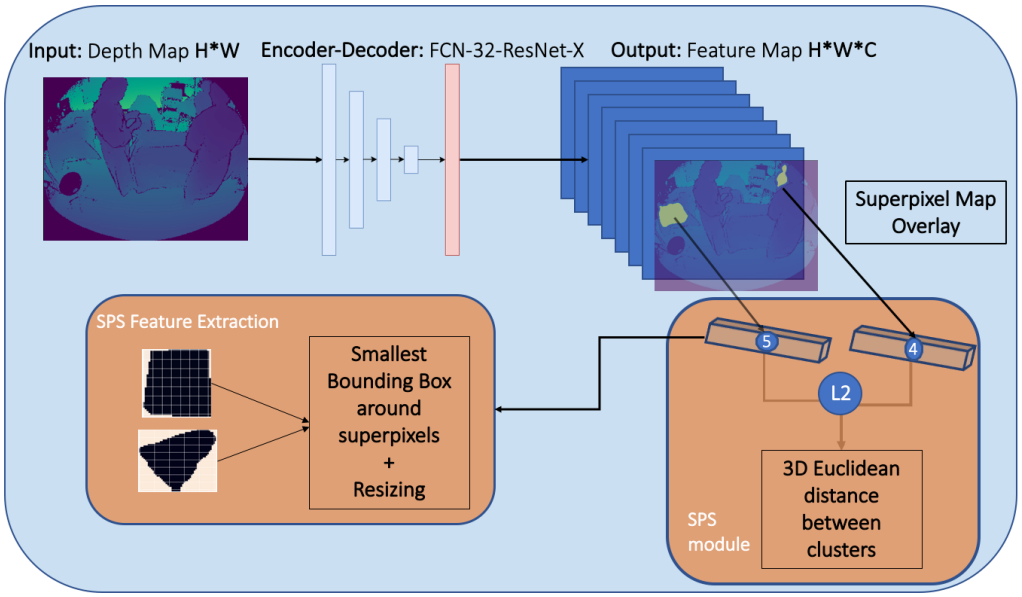

Figure 4.6: Framework used for self-supervised learning; numbers 4 and 5 on feature vectors refer to Figure A.3.

Architecture: To retrieve features at a superpixel level, we need to upscale our feature space. This work uses FCN-32 [Shelhamer 2014] with a ResNet backbone [He 2015] as our encoder-decoder architecture for simplicity and computational efficiency. This architecture only contains one deconvolutional layer and one upsampling module, resulting in a $224 \times 224 \times 32$ feature map. We keep the same architecture for our semantic segmentation experiments as for our pretext task. In contrast, we only keep the pretrained ResNet encoder with a global average pooling and a fully connected layer on top for our activity classification experiments.

Superpixel Sampling module (SPS): We sample the cluster features from the decoder output map. We use the superpixel map's external knowledge to retrieve the position from which we sample our features. For each pair of clusters, we consider the smallest bounding box around the superpixel and extract the features from those two bounding boxes (see Fig. 4.6). Once those features are pulled from the decoder output, we resize them to compute an element-wise l_2 loss between them. The superpixel features are resized to $20 \times 20 \times 32$.

4.2.4 Semi-supervised Learning: Semantic Segmentation & Activity Classification

Methodology: Some critical studies on self-supervised learning have demonstrated that results were dependent on the complexity of the dataset, of the downstream task at hand, of the architecture used, and, of course, of

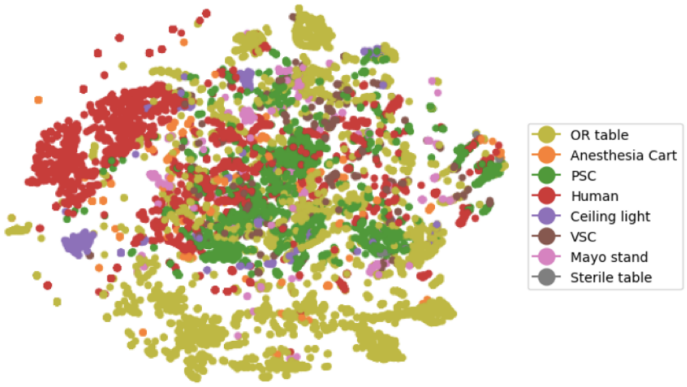

Figure 4.7: t-SNE [van der Maaten 2008] visualization of the superpixel features learned from the pretext task.

the amount of supervision [Newell 2020, Asano 2020]. Our aim is to answer those concerns by demonstrating the utility of our method on two different architectures with different complexities and two different downstream tasks.

Evaluation Metrics: We use the same evaluation metrics as in the related publications [Li 2020b, Schmidt 2021, Sharghi 2020], where mean average precision (mAP) and mean intersection over union (mIoU) are used to effectively compare the results.

4.3 Experiments and Results

4.3.1 Operating Room Awareness Datasets

To demonstrate the applicability of our method, we use the following two recent datasets captured from the OR and containing depth image data.

4.3.2 Unsupervised evaluation of self-supervised task

The t-SNE [van der Maaten 2008] method is a dimensionality reduction method widely used in computer vision to evaluate qualitatively the features learned by a neural network. In our case, each point in the 2D point cloud represents the features belonging to one superpixel and extracted from our SPS module.

Superpixel segmentation provides an oversegmentation of our image, so it is very relevant to cluster the features extracted from those regions. These features are obtained without supervision, and as we can see on Fig. 4.7, our pretext task manages to learn identifiable clusters for most of the semantic classes appearing in our semantic segmentation dataset. This is visible for less frequent classes like the ceiling light and more frequent classes like the human and OR table.

4.3.3 Semi-supervised learning and data efficiency experiments

Pretraining protocol: We evaluate our method against three baselines:

(1) In the first setting, we train from scratch without pretraining to measure the benefits of the other self-supervised methods.

(2) The first self-supervised baseline is RotNet [Gidaris 2018], which is trained for 200 epochs to predict different rotations that have been applied to the initial image following the implementation from *Gidaris et al.* [Gidaris 2018].

(3) The second self-supervised baseline is CPC v2 [van den Oord 2018] trained using the authors' implementation for 200 epochs.

Our pretext task is trained with a learning rate of 3e-4 and a batch size of 32 for 200 epochs. We also ensure that all our baselines are trained fairly by saving only the best-performing model on our validation dataset over 200 epochs.

Finetuning protocol: We evaluate our method semi-supervised with different amounts of annotated data (2%, 5%, 10%, 20%, 50%, 100%). We follow the usual semi-supervised learning protocol and run our experiments with ResNet-18 and ResNet-50 to show that our results do not depend on the network complexity. Our results are averaged across five different random splits for all different data regimes, to account for the randomness introduced by sampling a small amount of data, as done in [Roß 2018].

Downstream performance: The results are shown in Fig. 4.8. They demonstrate the usefulness of the proposed task as a new pretraining task, as it outperforms training from scratch, as expected. The gap becomes smaller for all self-supervised pretraining experiments as we gradually increase the amount of supervision. Our pretrained task consistently outperforms the two self-supervised approaches on the low-data regime across the two architectures and tasks. It performs similarly in the high-data regime for both tasks. The dominant takeaway is that self-supervised initialization is especially helpful in lower-data regimes and often retains a small advantage even at higher data fractions. For example, on activity classification, our proposed task achieves the same mAP performance as training from scratch using only half the number of annotations at 5% and 50%.

Statistical Significance Analysis: Using a Wilcoxon signed-rank test, we further measure the statistical significance of our proposed pretext task performance compared to the "scratch" baseline. We perform the significance analysis for all data fractions, based upon the collected p-values adjusted by Dunnett's test across splits and Bonferroni-Holm correction across data fractions. Our proposed method shows significant ($p << 0.05$) improvement on both downstream tasks for the low regime data, up to 20% for semantic segmentation and up to 10% for activity classification. It even shows significant

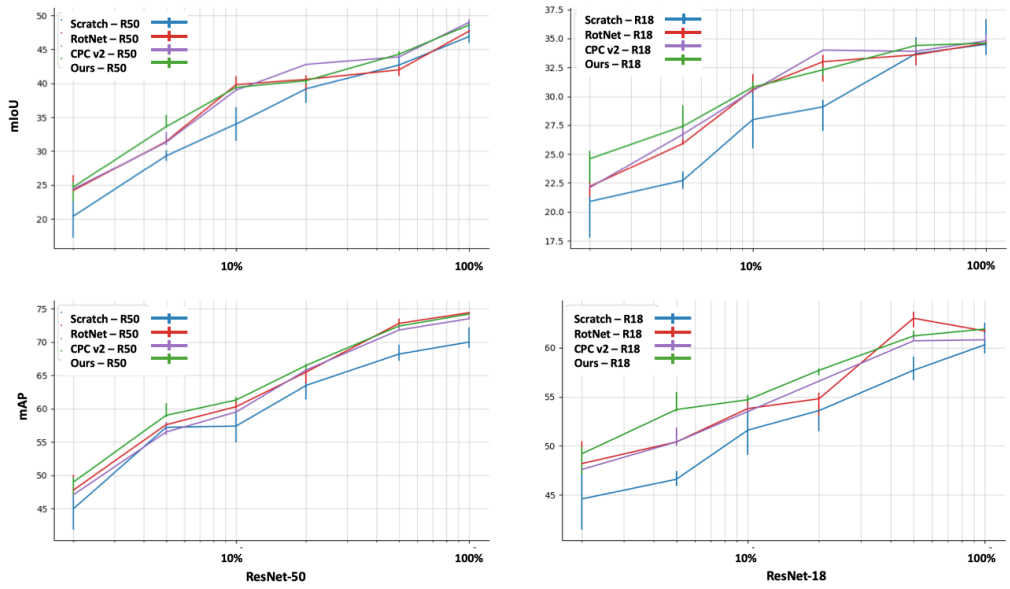

Figure 4.8: Median mIoU and mAP with Interquartile Range (IQR) as a function of training available labels as described in section 4.3. Our method outperforms the baseline method without pretraining and is on par with other self-supervised methods.

contributions for semantic segmentation to beat a narrower 0.01 significant level on the three lower data fractions (2%, 5%, 10%).

4.4 Conclusion

In this chapter, we explored the use of self-supervision on depth maps to improve semantic segmentation and activity recognition tasks in the surgical operating room. To this end, we introduced a novel 3D-based pretext task that leverages the geometric structure of the OR layout. Our approach was compared with established self-supervised methods such as RotNet [Gidaris 2018] and CPC v2 [van den Oord 2018], showing notable performance gains in low-data regimes, highlighting the effectiveness of our depth-driven supervision strategy.

While promising, our method currently falls short of state-of-the-art performance benchmarks [Li 2020a], as reported in Table 3.4. This is partly due to using a relatively simple FCN-32 architecture [Shelhamer 2014] and ResNet backbones [He 2015]. These choices were made to isolate the pretext task's effects, rather than optimize final segmentation performance. We leave it to future work to focus on integrating our self-supervised depth-based strategy into more advanced segmentation frameworks to realize its full potential.

While our self-supervised approach enables unsupervised oversegmentation of the scene, the resulting superpixels lack semantic content, as they are derived solely from depth cues. This limitation motivates the exploration of strategies that incorporate higher-level semantic features to enhance the integration of abstract modalities. Accordingly, the next chapter shifts focus toward incorporating object detection outputs, rich in both categorical and spatial information, to bridge this semantic gap and advance the effectiveness of our multimodal scene understanding framework.

5

Self-Supervised Masked Object Embedding Prediction for Object-Centric OR Activity Recognition

Contents

This chapter tackles the challenge of recognizing actions based on spatial layouts by employing specialized object and person detectors. It focuses on object-centric video analysis, modeling the geometric interactions among surgical devices to generate informative representations for clip-level action classification. The primary goal is to explore how an object-centric learning approach can enhance the recognition of surgical activities by utilizing the spatial positions of surgical instruments and clinicians as key distinguishing features.

To achieve this, our approach unfolds in two primary steps. First, we introduce a model that leverages category information from semantically defined surgical objects, alongside spatial details derived from bounding box coordinates. This initial model employs straightforward MLP modules for object-centric reasoning. We then integrate these object-centric representations with traditional features extracted through a 3D CNN [Carreira 2017].

In the second step, we further advance our exploration by transitioning to a permutation-invariant transformer-based architecture. This improved model uses object-level features within a masked autoencoding framework, establishing a pretraining objective designed to enhance the overall effectiveness of our surgical activity recognition pipeline. Data-efficiency experiments conducted throughout demonstrate the performance gains achieved through this two-step, object-centric methodology.

This chapter is adapted from the following publication:

> HAMOUD, I., JAMAL M.A, SRIVASTAV, V., MUTTER, D., PADOY, N., MOHARERI, O. (2023). ST(OR)2: SPATIO-TEMPORAL OBJECT LEVEL REASONING FOR ACTIVITY RECOGNITION IN THE OPERATING ROOM. MEDICAL IMAGING WITH DEEP LEARNING (MIDL)

5.1 Introduction

Recognizing surgical actions from low-resolution RGB or Time-of-Flight (ToF) videos presents significant challenges due to visual clutter from various surgical instruments and the complexity of multi-agent interactions. Recent research indicates that incorporating object-centric features can effectively address these challenges by highlighting critical visual regions related to surgical activities. Prior approaches, such as STIN [Materzynska 2019], STRG [Wang 2018], STLT [Radevski 2021], and ORViT [Herzig 2022], leveraged annotated bounding boxes or general-purpose object detectors, primarily targeting single-agent scenarios within standard environments. However, robotic surgical contexts require specialized, domain-specific detectors capable of accurately handling unique surgical instruments and interactions among multiple clinicians.

To this end, we initially introduce *ST(OR)*^{2}, a framework that utilizes multilayer perceptron models built on spatial and semantic object-centric

features. These features are derived from dedicated detectors trained to identify surgical devices and clinicians in Robotic-Assisted Surgery (RAS) environments. The extracted object-centric representations seamlessly integrate with conventional appearance-based features, enhancing overall activity recognition performance.

Building upon this, we propose an improved method, *ORDynaRe*, which advances the architecture of *ST(OR)*^{2} by introducing permutation invariance and effectively modeling multiple instances within the same object category, particularly for clinicians. ORDynaRe incorporates a self-supervised learning approach tailored explicitly for surgical action recognition. It begins with an off-the-shelf detector previously trained with supervised learning on annotated subsets of surgical data to identify clinicians and surgical instruments. Subsequently, this detector generates pseudo-bounding boxes for unannotated training data, enabling the creation of robust, object-centric representations. These representations are then projected into high-dimensional spaces, temporally aligned using bounding box confidence scores, and serve as inputs for a novel self-supervised pretraining strategy inspired by masked image modeling techniques (e.g., VMAE [Tong 2022], MaskFeat [Wei 2022], and MME [Sun 2022]). In contrast to traditional pixel-based masking, our approach applies masking directly at the object-token level within a transformer-based architecture, predicting both the classes and positions of masked tokens.

This self-supervised methodology substantially enhances ORDynaRe's generalizability, allowing it to achieve robust performance even with limited labeled data. Experimental evaluations demonstrate that ORDynaRe consistently improves object-centric model accuracy across varying supervision levels, matching or surpassing global, clip-based feature methods when trained with as little as 20% labeled data.

5.2 Methodology

In this section, we will detail the methodologies of the two different architectures we iteratively proposed to model spatio-temporal layout effectively for activity recognition in the OR. We first detail our MLP-based approach, which shows improvements, even though the design of the method does not allow for permutation invariance, and we were forced to aggregate the features coming from different entities of the same category into a single vector representation. Our extension of the architecture using multi-head attention [Vaswani 2017] allowed us to better handle different entities of the same category.

5.2.1 ST(OR)^2: A MLP Based approach

We propose Spatio-Temporal Object-level Reasoning in the Operating Room (ST(OR)^2) for OR surgical activity recognition. ST(OR)^2 takes as input the 2-

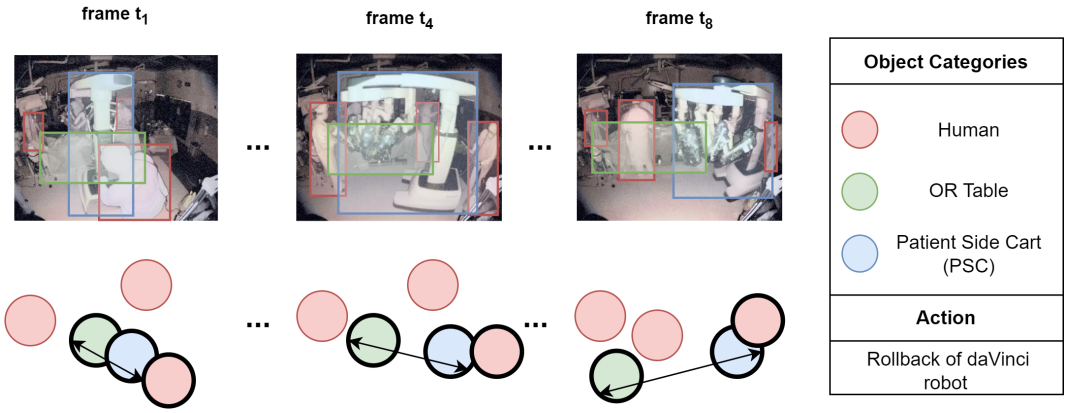

Figure 5.1: **Example video illustrating the "daVinci Rollback" action class.** Our method emphasizes geometric interactions between semantically identified objects. In this instance, the proximity of a clinician to the PSC and their movement away from the OR table serve as strong indicators for accurately predicting the action.

d bounding boxes extracted from a T frame-long clip. We use the bounding box geometric and semantic information about each object class to build our graph. We use our spatio-temporal object graph to reason over the objects' relative locations to recognize clip actions based on interaction dynamics. This backbone is the first stage of our long video segmentation, allowing us to extract reliable features.

In the second stage, we use the features extracted from our clip-based model to train a temporal sequence model to capture long-range dependencies in the videos.

Object-Centric Representation

We sample short T-frame-long clips from each phase of the longer OR surgical videos to train our backbone. Each object appearing in those frames will serve as a node of our graph. Inspired by STIN [Materzynska 2019], each node will be associated with specific features grounded on the position and category of the object (cf Figure5.2).

Person/Object Detection We infer bounding boxes for all videos using two Cascade Mask RCNN [Cai 2018] with a ConvNext backbone [Liu 2022] pretrained on the OR-Det dataset introduced in Chapter 3, containing OR images for human and RAS-specific objects. We will use N bounding boxes for each frame of the videos. If a frame has fewer than N boxes, the remaining feature vectors are padded with zeros.

Spatial Position Embedding We represent object position using their bounding box coordinates as a 4-d vector containing center coordinates and the width and height of the box. This 4-d vector is then forwarded to a

multilayer perceptron to obtain a d-dimensional embedding.

$$\sigma_{i,t} = MLP_{SPE}(box_{i,t}) \quad i \in \{1,\dots,N\},\ t \in \{1,\dots,T\} \tag{5.1}$$

Category Embedding We use the knowledge about the classes of objects to enhance each node's representation. Each of those C classes will be associated with a d-dimensional learnable embedding, which is randomly initialized from an independent multivariate normal distribution.

$$class_{i,t} = c \in \{1,\dots,C\} \quad \kappa_{i,t} = Embed_c \quad i \in \{1,\dots,N\},\ t \in \{1,\dots,T\} \tag{5.2}$$

Both spatial position and category embedding are concatenated and passed through an MLP to obtain a fused representation for each graph node.

$$x_{i,t} = MLP_{Fusion}(\sigma_{i,t}||\kappa_{i,t}) \quad i \in \{1,\dots,N\},\ t \in \{1,\dots,T\} \tag{5.3}$$

Spatio-Temporal Reasoning

Category-wise Aggregation We aggregate features of objects in the same category by summing them together. This alleviates any need to track different instances of the same class across frames, but it also discards instance-specific information for humans.

$$\varphi_{c,t} = \sum_{class_{i,t}=c} x_{i,t} \quad c \in \{1,\dots,C\},\ t \in \{1,\dots,T\} \tag{5.4}$$

Temporal-Category Interaction Module Using the aggregated features for each object category, we first carry out temporal reasoning across categories after concatenating the $\varphi_{c,t}$ features for each frame t of the clip.

$$\varphi_c = MLP_{Temp}(\varphi_{c,1}||...||\varphi_{c,T}) \quad c \in \{1,\dots,C\} \tag{5.5}$$

Once we obtain a feature vector representing the temporal evolution of each object category, we perform category-wise reasoning over the concatenated features of each category to get a clip-level representation. We use a cross-entropy loss on the output probabilities to train our clip classification backbone.

$$\phi_{clip} = MLP_{Category}(\varphi_{c=1}||...||\varphi_{c=C}) \tag{5.6}$$

Our object-level representation can also easily be combined with video appearance features extracted from a 3D CNN; in our experiments, we will be using I3D [Carreira 2017].

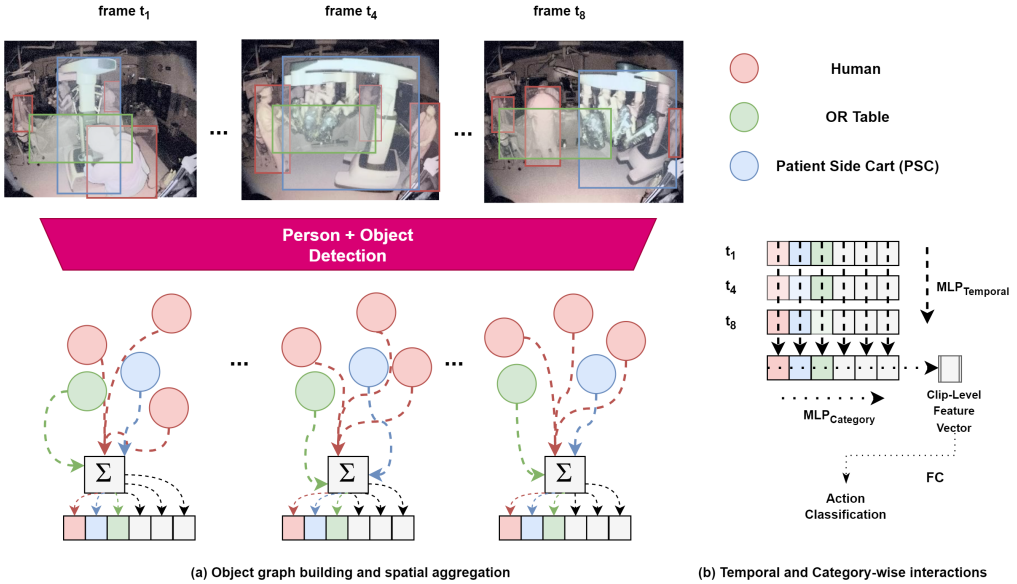

Figure 5.2: **Architecture of the ST(OR)^2 method:** We first build our object graph and aggregate features category-wise. In the second step, we reason over time for each category, and then we reason over categories to obtain a clip-level feature vector for action classification.

Temporal Sequence Modeling

Following our clip-based feature extraction, each video is then represented as $v_i = \{\phi_1, ..., \phi_T\}$ with ϕ_t being the feature extracted from the t^{th} clip. Those features can then be concatenated with the I3D features and fed to Uni-GRU [Chung 2014]. This allows us to deal with long-range temporal dependencies and obtain a more robust temporal segmentation for long videos.

5.2.2 ORDynaRe: A self-supervised Spatial-Temporal Transformer approach

Unlike CNNs and RNNs, which are designed with specific inductive biases based on the input type. Transformers [Vaswani 2017] can take as input any type of unordered input, as long as the dimensions of the input are kept constant across batches. They can be further expanded with specific spatio-temporal information in positional embeddings. Our unordered set of object layout information falls into this category. One way to alleviate the manual supervision for activity recognition is to pretrain the model in a self-supervised fashion on unlabeled videos before fine-tuning it for the downstream task. Below, we introduce the proposed object-centric representation, present our pretraining strategy, and discuss how it is integrated with a transformer-based architecture.

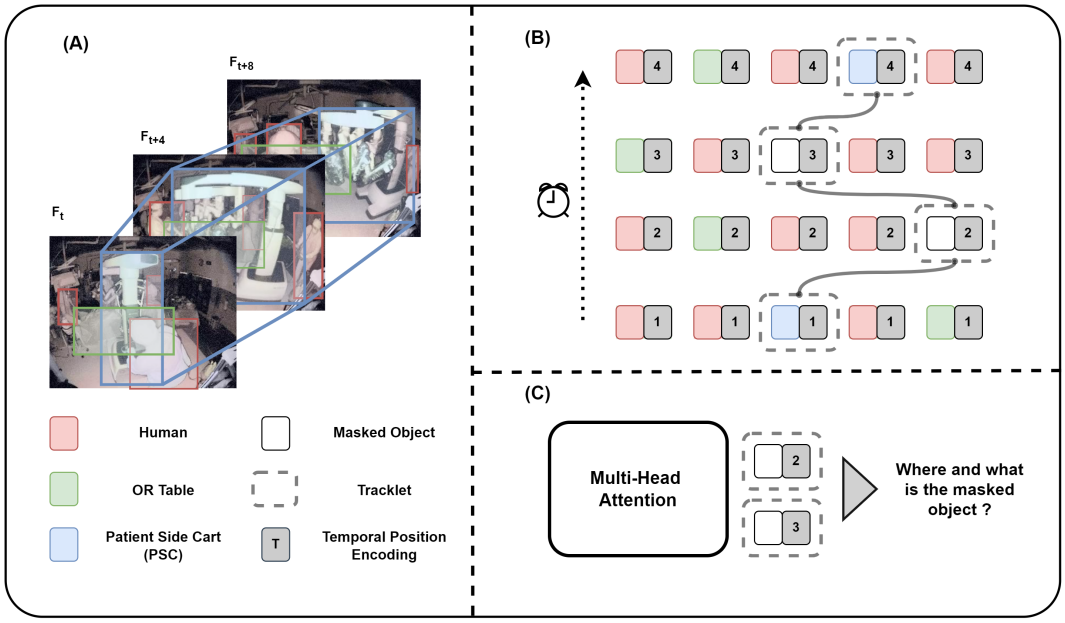

Figure 5.3: **A schematic of our masking strategy.** (A) Objects and persons are detected and tracked across video frames, and the longest tracklet is selected for masking. (B) The first and last bounding boxes of the tracklet are retained as context, while object features in intermediate frames are hidden. (C) A transformer processes the unmasked features to predict the position and semantic class of the masked objects.

Object-Centric Representation

Tracking of Human/OR Objects To segment each frame of our clips into object representations, we make use of two pretrained Cascade Mask RCNN [Cai 2018] object detectors using a ConvNext backbone [Liu 2022], same as in 5.2.1. The detectors are pretrained on OR images for human and RAS-specific objects. We use N bounding boxes per frame of the videos, with N being the maximum number of detected objects in a single frame. We later link the boxes temporally using heuristics on bounding box overlap and object detection scores similarity.

Object Token Representation Each detected object is represented by its (a) bounding box information $b_{i,t}^{(box)} \in \mathbb{R}^4$ that is fed into a multilayer perception (MLP) to obtain the spatial position embedding $\sigma_{i,t} \in \mathbb{R}^d$, (b) category embedding $\kappa_{i,t} \in \mathbb{R}^d$ initialized from a multivariate normal distribution for each of the C object categories, and (c) ROI pooled features $\rho_{i,t} \in \mathbb{R}^d$ from the detector. The spatial position embedding and category embedding are fused and then concatenated with the ROI features, unlike *ST(OR)*^{2}. We further add a temporal positional encoding $PE_t \in \mathbb{R}^{2\times d}$ to each object token $\phi_{i,t}$ based on the order of the frame it belongs to in the clip, as can be

seen in Figure 5.4. As transformers [Vaswani 2017] are order-invariant, we need a way of including temporal information, which was handled by the temporal interaction module in *ST(OR)*^{2}.

$$\phi_{i,t} = [MLP_{fusion}([\sigma_{i,t}||\kappa_{i,t}])||\rho_{i,t}] + PE_t \tag{5.7}$$

Masked Object Prediction

Our unsupervised task follows the mask-and-predict paradigm of the existing masked video modeling task [Tong 2022, Wu 2021]. Still, it replaces the patch-level reconstruction objective by predicting the position and semantic class of the masked object.

Notations: In our case, we mask object embeddings and train the model to infer the content of their representations using its knowledge of unmasked objects. Specifically, during training, we multiply each object's representation by $\mu_{i,t} \in \{0, 1\}$ before feeding it into the transformer. Let $\phi'_{i,t}$ be the transformed token representation; we define $h^{(box)}$, $h^{(class)}$, $h^{(feature)}$ as the linear heads on top of each transformed token to predict respectively the bounding box coordinates, class and initial feature vector of the masked token.

Loss Functions: The unsupervised prediction loss is defined as the weighted sum of a cross-entropy loss $l^{(class)}$ on the semantics of each masked object, a regression loss $l^{(box)}$ to predict the bounding box coordinates, and an auxiliary loss $l^{(feature)}$ to regress the initial features using an l_2 loss. The "feature" loss serves as a regularization constraint since, in early training steps, the spatial position embedding of each object token has not been properly learned yet.

$$l_{i,t}^{(mask)} = l_{i,t}^{(class)} + l_{i,t}^{(box)} + \lambda_{feature} \cdot l_{i,t}^{(feature)} \tag{5.8}$$

$$l^{(mask)} = \sum_{i,t} (1 - \mu_{i,t}) \cdot l_{i,t}^{(mask)} \tag{5.9}$$

Similarly to [Carion 2020b], we adopt a soft version of Intersection over Union in our box loss, together with an l_1 loss; this helps predict more accurate bounding boxes and deal with frequent scale issues.

$$l_{i,t}^{(box)} = \lambda_{L1} \cdot ||h^{(box)}(\phi'_{i,t}) - b_{i,t}^{(box)})||_1 + \lambda_{iou} \cdot (1 - gIoU(h^{(box)}(\phi'_{i,t}), b_{i,t}^{(box)})) \tag{5.10}$$

Masked Object Sampling: For each clip we consider every short-term tracklets of objects $\{(t, box_{j,t}) | \tau_i(t) = j,\ \forall t, i\}$ as a masking candidate. Ideally, we would like our sampling strategy to ensure class balance in terms of masked instances to ensure that our model learns meaningful information

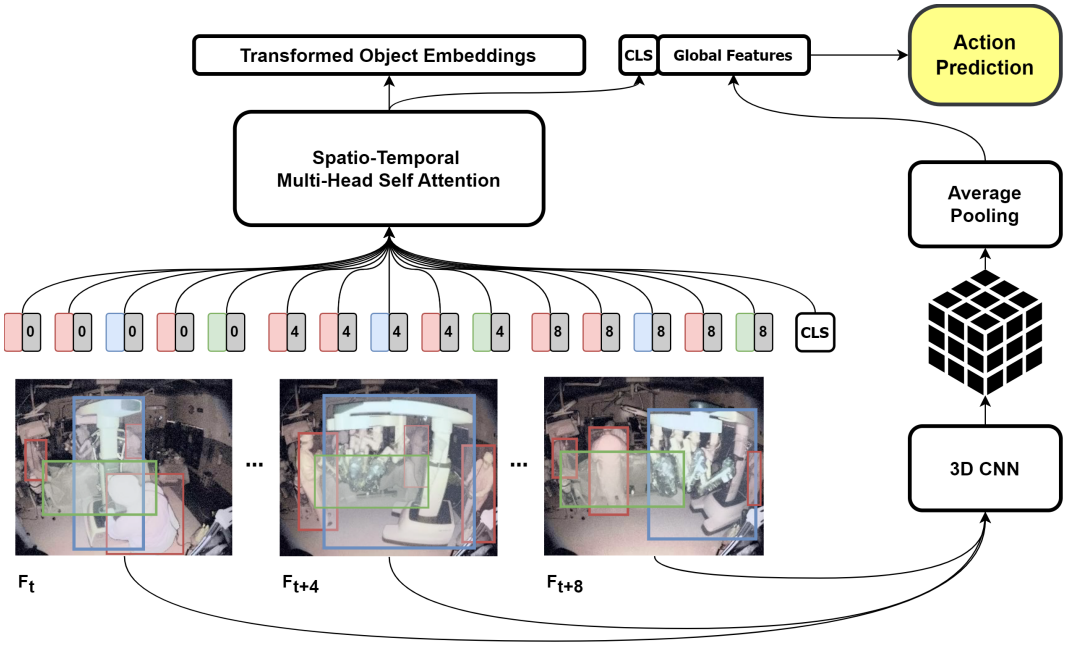

Figure 5.4: **Architecture of ORDynaRe for action recognition:** We first tokenize our videos at the object level. A special $[CLS]$ token is appended to the object tokens. We reason over space and time jointly as our transformer attends to all objects across the video. The $[CLS]$ token can be fused with the appearance features using late concatenation for action prediction.

for all classes. We randomly sample a category and pick a random tracklet belonging to this category; if it is present, we will otherwise pick the longest tracklet candidate. We mask the object tokens corresponding to this tracklet, but keep the first and last token as context (Figure 5.3). Of course, other settings using multiple objects per frame masking could have been considered. The current setting discards the potential benefits of non-local prediction by using longer temporal ranges.

Fine-Tuning on Action Recognition

We propose a simple yet effective transformer-based approach operating on a finite sequence of object vectors in $\mathbb{R}^d$. We append a learned vector $[CLS]$ (Figure 5.8) as the first token of each sequence (akin to the $[CLS]$ special token in [Bao 2021]) and use the output vector corresponding to that position for clip-level action recognition. This vector can later be concatenated with the globally extracted feature vector from an appearance model. We use a linear head $h^{(act)}$ on top of the $[CLS]$ token to output a probability vector.

When fine-tuning our model, we initialize the transformer and the linear layers that define our object tokens with the self-supervised pretrained weights.

5.3 Experiments and Results

5.3.1 Dataset

We use the OR-AR dataset [Sharghi 2020], a large-scale dataset containing up to 400 full-length videos from 103 surgical procedures. The collection of the videos is achieved using 4 Time-of-Flight cameras positioned in key placements on two different carts to capture the full OR comprehensively.

Temporal Annotations: Nine activities relative to RAS are annotated on the dataset. The activity classes are highly imbalanced. Rollback and Roll-out of the daVinci usually last between one and two minutes, while activities like Patient Preparation can last more than an hour. We use temporal annotations to fine-tune our method on the action recognition end task.

Spatial Annotations:

The human and object detectors we use to extract object information are trained on the OR-Det dataset introduced in Chapter 3. We provide the detailed performance per class of the object and human detectors in Table 5.5.

5.3.2 Experiments

In our ORDynaRe experiments, we use a multi-head Attention transformer [Vaswani 2017] on top of our learned object token embeddings. We set the number of layers and heads of our transformer to 8 and 12, respectively. The dimension d of the object embeddings is chosen as 512. We select I3D [Carreira 2017] as our feature extractor for our experiments involving appearance-based models, but our approach can be applied to any other spatiotemporal models. We measure performance on both clip classification and long-video phase recognition, which we refer to as surgical action recognition and surgical activity recognition, respectively.

Baselines

We examine the performance of our object-centric methods against multiple baselines: (1)*ORDynaRe (No SSL)* method serves as a fundamental baseline to demonstrate the utility of our pretraining. (2)*I3D* [Carreira 2017] is a popular 3D CNN pretrained on Kinetics. It has been largely adopted in the community. (3)*I3D+STRG* [Wang 2018] is an object-centric approach using box information output by a pretrained RPN [Ren 2016]. This showcases the effectiveness of our object-centric architecture. (4) TimesFormer [Bertasius 2021], and (5) MotionFormer [Patrick 2021] as the global image-centric baselines. We choose (6) STRG [Wang 2018] and (7) ORViT [Herzig 2022] as the local object-centric baselines. And also our (8-9) ST(OR)^2

Table 5.1: Results and comparison against baselines for OR surgical activity recognition on complete procedures. mAP (%) is reported across three splits for all data fractions, along with the average mAP

Surgical Activity Recognition							
Temporal Model	**Backbone**	**Visual Features**	**2%**	**5%**	**10%**	**20%**	**100%**
Uni-GRU	I3D	✓	19.7±2.8	39.2±1.9	53.5±1.5	79.5±0.9	90.7±0.6
	ST(OR)^2	✗	27.3±2.1	48.8±1.7	58.2±1.9	68.3±1.6	73.6±1.3
	I3D + ST(OR)^2	✓	**29.5±2.3**	**54.2±1.7**	**60.1±1.6**	**82.3±1.4**	**91.8±1.0**

Unsupervised pretraining.

We pretrain our model defined above by minimizing the previously described $l^{(mask)}$ loss, using a stochastic gradient descent as our optimizer. We use a cosine learning rate schedule over 30 epochs, with 0.01 as the initial learning rate and a weight decay of 10^{-5}. We use this pretraining better to initialize our object-centric model on the end task.

Fine-tuning Experiments

Utility of pretraining We show that our pretraining leads to substantial gains in performance. As shown in Figure 5.5, the self-supervised pretrained ORDynaRe consistently overperforms the one trained from scratch across all levels of supervision. Our self-supervised model overperforms or is on par with the I3D baseline for up to 20% of labeled data. This highlights the utility of pretraining our object-centric model.

Combining Object-Centric and Appearance Features When combined with the I3D global features, the performance of our method is further improved. The improvement introduced by our approach is noticeable in both surgical clip action recognition Figure 5.5, Table 5.2, and long-video activity recognition, as shown in Table 5.3. It further emphasizes exploiting object information priors combined with appearance-based visual features.

Furthermore, the features learned from our pretrained object-centric approach lead to an impressive 90.47% mAP when trained with Bi-GRU on surgical activity recognition without using any appearance-based model.

5.4 Conclusion

This chapter introduced a novel, geometrically grounded, object-centric approach to surgical video understanding, focusing on activity recognition in the OR. Our method constructs a spatiotemporal graph based on the geometric layout of clinicians and surgical instruments, enabling robust reasoning about surgical activities. It operates solely on RGB images without requiring camera calibration or point cloud data, making it significantly more memory and data-efficient than existing 3D-based methods.

We proposed a self-supervised pretraining strategy based on masked

Table 5.2: Results and comparison against baselines for clip-based surgical action classification. top-1 accuracy (%) is reported.

Surgical Action Classification		
Backbone	**Object-Centric Model**	**Top-1 Accuracy**
I3D	×	87.9±0.9
TimesFormer	×	85.7±0.4
MotionFormer	×	84.5±1.2
I3D	STRG	88.4±1.2
MotionFormer	ORViT	84.1±1.4
×	ST(OR)^2	47.3±2.1
I3D	ST(OR)^2	89.4±0.8
×	ORDynaRe	57.7±1.2
×	ORDynare (SSL)	60.6±0.7
I3D	ORDynare (SSL)	**91.4±0.9**

Table 5.3: Results and comparison against baselines for OR surgical activity recognition on complete procedures. We report mAP, Precision, Accuracy, and Recall.

Surgical Activity Recognition					
Temporal Model	**Method**	**mAP**	**Precision**	**Accuracy**	**Recall**
Uni-GRU	I3D	89.62	87.12	92.51	82.78
	ORDynaRe	80.53	71.84	87.37	70.17
	I3D + ORDynaRe	**91.13**	**89.16**	**95.22**	**85.41**
Bi-GRU	I3D	93.45	90.74	96.22	85.64
	ORDynaRe	90.47	81.22	89.13	78.27
	I3D + ORDynaRe	**93.92**	**91.53**	**96.88**	**86.02**

Table 5.4: Hyperparameter table for both self-supervised pretraining and fine-tuning.

Hyperparameters	**pretraining**	**Fine-tuning**
MLP hidden size	256	256
Nb. of object tokens/frame	15	15
Nb. of frame/clip	8	8
Sampling	1 fps	1 fps
Dropout	0.15	0.15
Learning rate	3×10^{-3}	6×10^{-3}
Warmup Epochs	0	10

Table 5.5: Class-specific performance for object detection class-specific. mAP at IoU:0.5:0.95 (%) is reported on the testing set for our detector.

Object Detection Performance						
Method	**Human**	**Table**	**Gurney**	**PSC**	**OR Table**	**VSC**
Cascade MaskRCNN	79.3	65.4	57.4	70.2	46.4	69.7

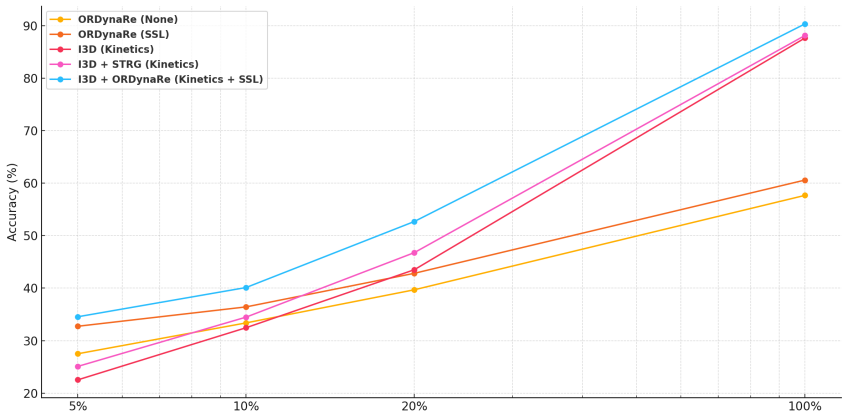

Figure 5.5: Results and comparison against baselines for ORDynaRe surgical action recognition on clip classification. Accuracy (%) is reported across two different seeds for all data fractions.

human motion prediction to enhance temporal modeling and object awareness. This strategy allows the model to learn from minimal context, effectively capturing clinician interactions and cross-view temporal consistency in an unsupervised manner.

Unlike prior methods that rely on dense point cloud inputs, our approach induces object-centric representations using only coarse geometric cues such as bounding boxes. While already effective with limited labeled data, future work could incorporate richer spatial features like human pose estimation to improve activity recognition performance.

Overall, this contribution demonstrates the effectiveness of combining geometric priors with self-supervised learning for scalable, efficient surgical video understanding, paving the way for more accessible and robust models in data-scarce clinical environments.

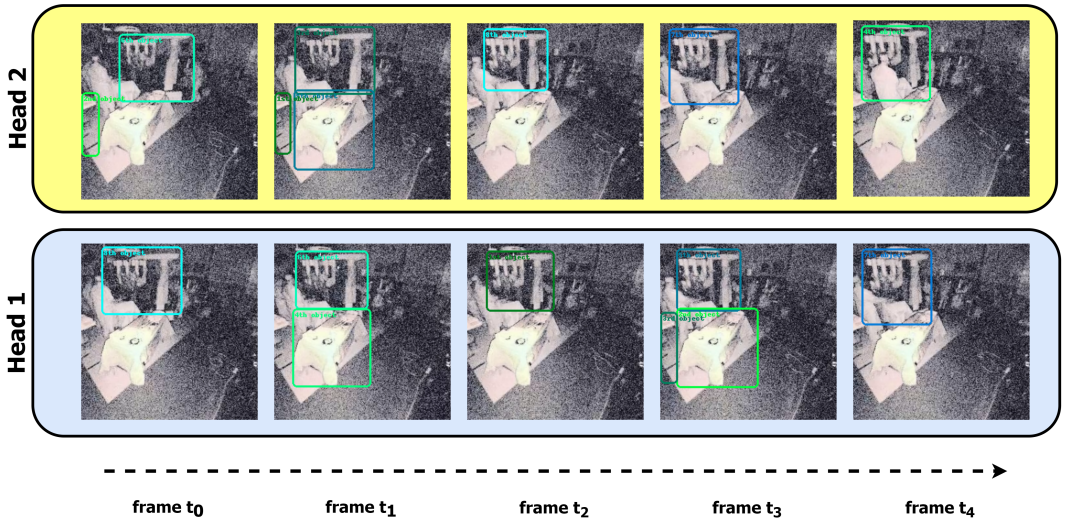

Figure 5.6: We highlight the eight objects with the highest attention score with the [CLS] token across three different heads of the last layer. The action conducted in the clip is the Roll-up of the daVinci. The representation of the PSC across different heads illustrates its importance in recognizing this specific action.

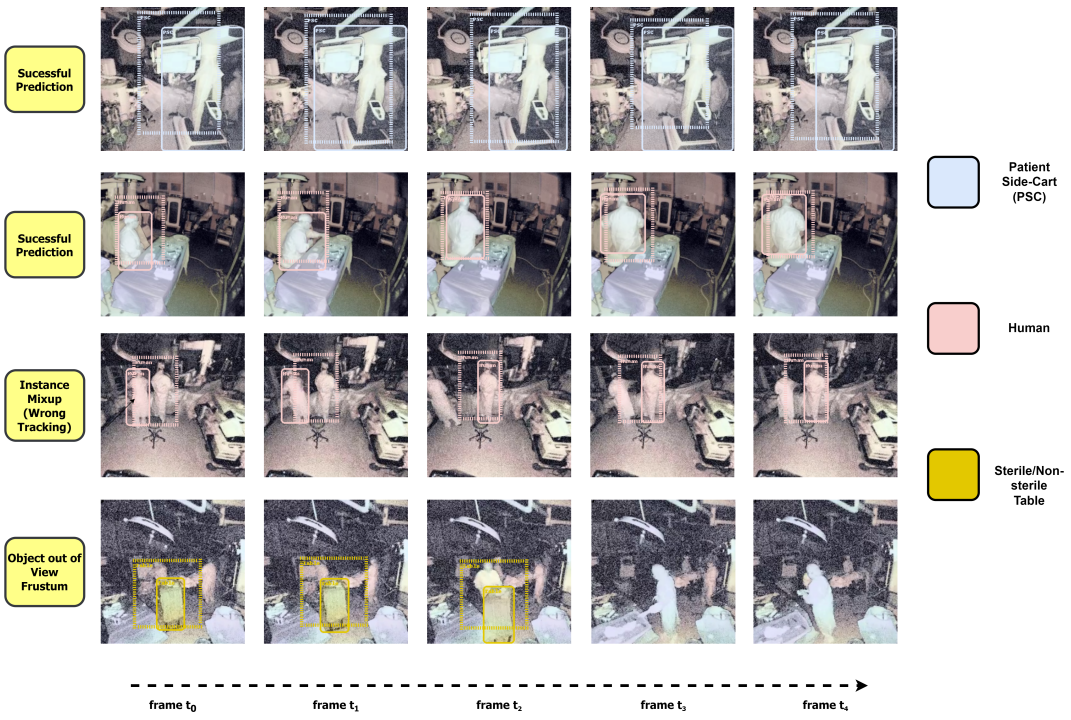

Figure 5.7: Box and class predictions on four different clips. The dashed bounding box is the predicted bounding box while the solid line is the ground-truth bounding box.

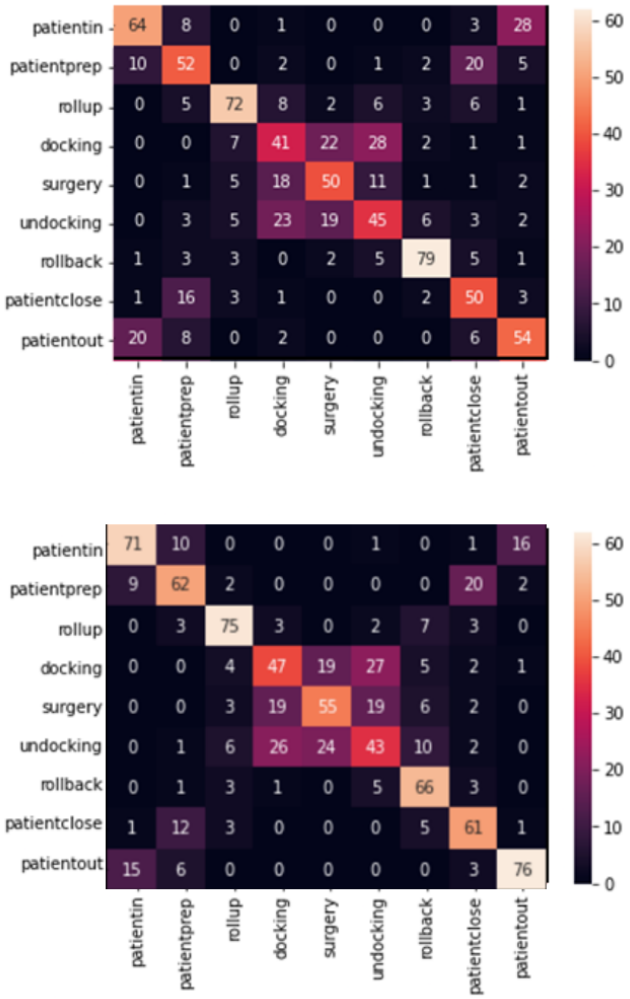

Figure 5.8: Comparison of normalized confusion matrices for action recognition with (top) and without (bottom) self-supervised pretraining.

6

Robust OR Multiview Activity Recognition via Self-Supervised Multimodal Feature Alignment of Video and Pose

Contents

This chapter presents a novel calibration-free multi-view, multi-modal pretraining framework to enhance surgical activity recognition (SAR) by effectively aligning human pose representations with visual data from uncalibrated camera setups. To address the challenges posed by continuous 2D pose data, we employ Pose as Compositional Tokens (PCT) [Geng 2023], which discretizes pose information into structured tokens, facilitating integration within a dual-encoder architecture. Our framework introduces specialized pretraining objectives, including view invariance, geometric constraints, and masked pose modeling, to align pose and visual embeddings across multiple views robustly. Extensive evaluations demonstrate that our approach significantly outperforms existing methods, particularly in single-view deployment scenarios, thereby advancing the capabilities of SAR systems in complex surgical environments.

This chapter is adapted from:

HAMOUD, I., SRIVASTAV, V., JAMAL, M. A., MUTTER, D., MOHARERI, O., PADOY, N. (2025). MULTI-VIEW VIDEO-POSE PRETRAINING FOR OPERATING ROOM SURGICAL ACTIVITY RECOGNITION. ARXIV PREPRINT ARXIV:2502.13883.

6.1 Introduction

The modern Operating Room (OR) is a high-stakes, fast-paced sociotechnical environment where clinicians work collaboratively to ensure safe and efficient surgical procedures. To support these efforts, ORs are increasingly equipped with advanced sensors, including external cameras, to monitor and analyze clinical activities. By leveraging these sensor-enhanced capabilities of the OR, context-aware systems have emerged as a promising tool to optimize clinical workflows, support intra-operative decision-making, and enable early detection of adverse events through automated analysis of clinical processes [Maier-Hein 2022]. Recent developments in OR applications highlight this potential, including radiation risk monitoring in hybrid surgeries [Rodas 2018, Ladikos 2010], surgical workflow recognition [Padoy 2008, Czempiel 2020], and semantic scene understanding [Murali 2022, cCaughan Koksal 2024].

A key component of such systems is *surgical activity recognition* (SAR), which aims to detect different activities or phases in long untrimmed videos recorded from external multi-view cameras. Recent SAR models [Zhang 2021, Twinanda 2016, Sharghi 2020], inspired by advances in action recognition, use clip-based approaches to segment videos into temporal phases. However, these approaches do not fully exploit the multi-view knowledge from the multi-camera setups and mainly rely on clip-level or global image features, overlooking fine-grained details of clinicians' movements. Some recently proposed methods based on the 4D-OR dataset

address this limitation but require calibrated multi-view camera systems and advanced point-cloud processing for semantic scene graph generation, which is then used for surgical activity recognition [Özsoy 2023]. However, these methods can be computationally expensive and rely on calibrated multi-view camera setups. These are challenging to acquire in practical OR settings, especially in robot-assisted surgical procedures where vision cameras are mounted on the movable surgical robot [Sharghi 2020].

As clinicians are the main dynamic actors in the OR, their fine-grained localization is crucial for reliable SAR systems. Human pose estimation, a computer vision task that localizes 2D body keypoints, has started to work remarkably well even in complex scenarios [Cao 2017, Geng 2023]. SAR models can significantly improve activity recognition accuracy by explicitly integrating fine-grained pose information.

In parallel, computer vision has witnessed significant advances in multimodal pretraining [Radford 2021, Li 2021, Jia 2021], a paradigm that bridges vision and language modalities. Models like CLIP [Radford 2021] and ALIGN [Jia 2021] have demonstrated the ability to learn generalized multimodal representations by aligning visual concepts with natural language descriptions using large-scale paired image-caption datasets. These models have enabled a shift from task-specific to more generalist models in a unified framework capable of handling diverse downstream tasks [Zou 2024, Lin 2023].

Motivated by these developments, this work introduces and investigates a key research question: how can human pose representations be effectively aligned with *uncalibrated* multi-view camera images in a *multi-view multimodal pretraining* framework? By addressing this question, we aim to improve the performance of SAR systems as a downstream task by leveraging human pose estimation, multi-modal pretraining, and multi-view video analysis.

However, the task is non-trivial and presents challenges regarding suitable architecture design and effective pretraining objectives. From an architectural perspective, we propose a dual-encoder that processes both vision and human pose modalities, similar to common vision-language architectures [Radford 2021]. However, unlike vision-language architectures where text is a *discrete* modality - with words or subwords transformed into discrete token representations - the human pose is typically represented as *continuous* 2D keypoints. To overcome this challenge, we propose to use the *Pose as Compositional Tokens* [Geng 2023], which tokenizes the continuous 2D human pose coordinates into discrete tokens. These tokenized embeddings convert the continuous poses into discrete tokens and handle occlusions well by leveraging the dependency between joints encoded in the discrete pose tokens.

Building on this architecture, we design a set of *pretraining objectives* to

align pose and vision embeddings while exploiting the multi-view context. The pretraining objectives follow the concept of CLIP [Radford 2021], where discrete pose embeddings are brought closer to the corresponding view's image embeddings, and embeddings of different images are pushed apart using InfoNCE loss [van den Oord 2018]. In the multi-view setting, we propose to extend further the idea to achieve *view invariance*: the model not only brings the pose embedding closer to its corresponding camera view but also aligns it with embeddings from other camera views at the same time stamp, and vice versa.

While these constraints help align multi-view pose and vision embeddings, they may still lack geometric alignment, leading to suboptimal downstream performance. To address this, we propose two additional geometric constraints to improve representation quality: *cross-modality constraints* - these constraints ensure that pose and visual embeddings are geometrically consistent across modalities, and *in-modality constraints* - these constraints enforce consistency within the pose or visual modality itself, enhancing structural coherence, similar to [Goel 2022].

Finally, we also leverage *masked modeling*, a technique widely used in visual and language representation learning [Tong 2022, Sun 2022, Wei 2022]. In masked image modeling, a portion of an image is hidden, and the model learns to predict the masked content based on its surroundings. Instead of applying this at the pixel level, we extend the idea to *pose tokens*. Specifically, we mask a subset of pose tokens and feed them into a transformer-based backbone, which learns to output a feature representation of the masked content. These representations are then input to a transformer decoder to predict the missing pose coordinates, encouraging the model to learn a robust representation of pose information.

In summary, this work introduces a novel *multi-view, multi-modal pretraining framework* by incorporating pose as compositional tokens, aligning embeddings across uncalibrated camera views, enforcing geometric constraints, and leveraging masked pose token prediction. We evaluate our framework on the SAR downstream task, conducting extensive ablation studies to analyze the contributions of each component and their impact on overall performance. A key highlight of our approach is its adaptability: even when fine-tuned with a single modality, our multi-modal pretraining framework achieves significant performance gains. Overall, we achieve significantly better results against strong baselines, thereby pushing the boundaries of surgical activity recognition, enabling a more accurate and reliable understanding of clinical workflows in calibration-free multi-view camera setups. Especially in single-view control experiments. This can be useful for a practical scenario where the activity recognition system should operate on a single camera at deployment time.

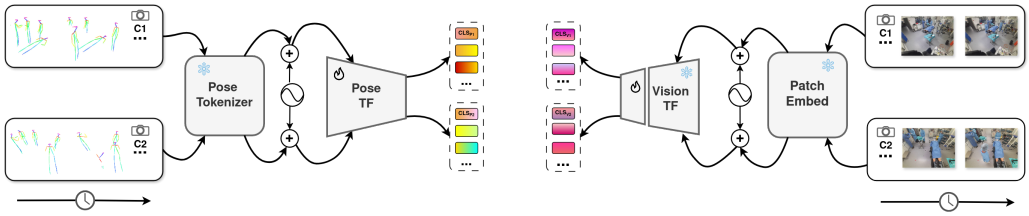

Figure 6.1: **Overview of our framework:** Given a video clip, we first extract all human poses using ViTPose-Base [Xu 2022]. We tokenize the poses using PCT [Geng 2023] and use a two-stream approach with MaskFeat [Wei 2022] on the vision features.

6.2 Methodology

In this section, we introduce *PreViPS*, our calibration-free multi-view multi-modal pretraining framework for surgical activity recognition (SAR). We first introduce the problem and then describe the *dual-encoder architecture*, detailing the *video encoder* for extracting visual features and our novel *pose encoder*, which converts continuous 2D human pose coordinates into discrete embeddings. Next, we describe our three multi-view multi-modal pretraining objectives, which align video and pose embeddings across camera views by enforcing *cross- and in-modality geometric constraints* and leveraging *masked pose token prediction*. Finally, we explain the *model finetuning process* for downstream SAR tasks, optimizing the learned representations for both multi-view and single-view surgical activity recognition. Through these architectural and training choices, *PreViPS* enables robust and efficient activity recognition in complex surgical environments.

6.2.1 Problem Overview

Given a training dataset of multi-view human-centric video clips $\mathcal{D} = \{\mathbf{x}|\mathbf{y}^*\}$ where $\mathbf{x} \in \mathbb{R}^{C\times T\times 3\times H\times W}$ is a multi-view video-clip set captured by C cameras over T frames with resolution $H \times W$, and $\mathbf{y}^* \in \mathbb{R}^{C\times T\times N_p\times 2\times N_j}$ represents the pseudo 2d human poses for N_p persons with $N_j(=17)$ number of joints (body keypoints) generated by an off-the-shelf human pose estimator, the goal to learn a joint latent space that correlates semantically similar video clips with the corresponding poses across camera views and vice versa.

Formally, our goal is to learn two mappings: $\mathcal{F} : \mathbf{x} \rightarrow \mathbb{R}^{C\times D}$ and $\mathcal{G} : \mathbf{y}^* \rightarrow \mathbb{R}^{C\times D}$, which map a video clip and 2d pseudo poses into a $C \times D$ dimensional latent vector, where D represents the embedding dimension. To learn these two mappings, we employ a dual-branch model with a vision branch $\mathcal{F}$ and a pose branch $\mathcal{G}$ using transformer-based *dual-encoder architecture*, described as follows.

6.2.2 Dual-encoder Architecture

Video Encoder

We employ *MaskFeat* [Wei 2022] as our video encoder. Given a multi-view video clip from C camera viewpoints, the encoder first applies a *patch embedding* layer, which employs convolution and linear projection to transform the video clip into a sequence of tokens. These token sequences are then processed by a Vision Transformer to produce contextualized video embeddings $I_c \in \mathbb{R}^D$ for each camera $c \in [1, \ldots, C]$, where I_c corresponds to the special [CLS] token, referred to as CLS_I^c, used in Vision Transformers.

Pose Encoder

Pose Token Representation Given a video clip, we use the VitPose-B [Xu 2022] as an off-the-shelf pose estimator to generate pose sequences for each camera view. We also gather identity information for each detected pose using the established SORT [Bewley 2016] algorithm.

Let $p_{i,t}^c \in \mathbb{R}^{2\times N_j}$ be the acquired pose coordinates at camera viewpoint $c \in [1, ..., C]$ for a person i and timestep t. To obtain a compact and meaningful representation for each single human pose, we pass $p_{i,t}^c$ through a frozen pose tokenizer [Geng 2023] to generate the following bottleneck representation: $\pi_{i,t}^c \in \mathbb{R}^D$. This structured representation models the dependency between body joints and provides a distinct discrete representation similar to the text modality in vision-language pretraining. For each camera stream, we also append a learned vector [CLS] as the first token of each sequence and use the output vector corresponding to that position for clip-level action recognition. More specifically, each camera stream pose latent representation is represented as,

$$\mathcal{Y}^c = \{CLS_J^c, \pi_{1,1}^c, ..., \pi_{N_p,T}^c\} \in \mathbb{R}^{D\times(N_p\times T+1)} \tag{6.1}$$

In this notation, $N_p = 8$ is the maximum number of detected persons per frame. To ensure a consistent number of inputs, if the number of clinicians in a frame is less than N_p, we pad the sequence with a special PAD token.

Positional Embeddings To encode spatiotemporal information in the pose sequences, we incorporate positional embeddings for various attributes such as time, track ID for persons, and viewpoint ID.

Concerning viewpoints, we adopt the method proposed by Geng et al. [Geng 2023], which involves introducing learnable 1D parameters that represent each viewpoint and timestep. For time and track ID, we utilize 2D sine and cosine functions as a form of positional encoding. These parameters are added to the features of each video pose token captured from different perspectives.

Network Architecture Given the previously defined representation, we adopt a vanilla transformer [Vaswani 2017] as the backbone network. The pose embeddings (aggregated with positional embeddings) described above are fed to the pose transformer $\mathcal{M}$.

$$\widehat{\mathcal{Y}} = \mathcal{M}(\Theta, \mathcal{Y}) \tag{6.2}$$

Here, Θ is the model parameters, and $\widehat{\mathcal{Y}}$ is the updated latent representation for pose information. The pose transformer comprises a stack of $L = 6$ multihead self-attention layers. Each layer in the pose transformer $\mathcal{M}$ has a standard architecture consisting of multi-head self-attention modules and feed-forward networks. The encoder outputs a sequence of pose embeddings of dimension D.

6.2.3 Aligning video and pose embeddings

In this section, we outline our approach to cross-modal alignment. Our contrastive objective is to optimize the encoders specific to each modality. These encoders map the global embeddings from each modality and viewpoint to ensure their representations are closely aligned. Let $I^c = CLS_I^c$ represent the learned embeddings for the video modality, and $J^c = CLS_J^c$ represent the learned embeddings for the pose modality, both at each camera viewpoint c.

Multi-modal Contrastive Learning

Cross-Modality Alignment Let us first address the cross-modal alignment between image and pose modalities. We want embeddings of the same samples from two viewpoints to be close to each other. Thus for each pair of camera views $(p, q) \in [1, ..., C]^2$, we aim to bring I_n^p and J_n^q closer together while pushing apart the other embeddings from the remaining samples in the minibatch of batch size N.

$$\mathcal{L}_{I/J} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{(p,q)\leq C}\log\left(\frac{\exp(\langle I_n^p, J_n^q\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle I_k^p, J_k^q\rangle/\tau)}\right) \tag{6.3}$$

Here τ is the temperature hyper-parameter that regulates the penalty to the hard negative samples.

In-Modality Alignment Similar to cross-modal alignment, we also propose objectives to increase the similarity of embeddings from the same modality that come from different viewpoints.

$$\mathcal{L}_{I/I} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{(p,q)\leq C}\log\left(\frac{\exp(\langle I_n^p, I_n^q\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle I_k^p, I_k^q\rangle/\tau)}\right) \tag{6.4}$$

We can define the $\mathcal{L}_{J/I}$ and $\mathcal{L}_{J/J}$ losses reflexively by adjusting the embeddings in the loss accordingly. The multi-modal contrastive objective in

PreViPS aims to align the video and pose representations by minimizing the loss function L_{Con} defined below:

$$\mathcal{L}_{Con} = \frac{1}{4}(\mathcal{L}_{I/J} + \mathcal{L}_{J/I} + \mathcal{L}_{I/I} + \mathcal{L}_{J/J}) \tag{6.5}$$

We refer to the pretraining following this loss, $\mathcal{L}_{Con}$, as CLIP* as it is an adaptation of the CLIP objectives to the video-pose modalities with multi-view constraints.

Sampling Policy Equations (6.3) and (6.4), are computed over N training instances, each in the form of a video-clip pose pair. A naive sampling policy may randomly sample instances from adjacent temporal segments, leading to semantically similar negative samples. This may confuse the model and hurt the final downstream performance. Therefore, we force each instance of our minibatch to be temporally distant. We divide the complete video into N segments, where N is the batch size, and sample one instance from each segment.

Geometric Consistency

We propose incorporating geometric constraints into our pretraining objectives, similar to CyCLIP in vision-language pretraining [Goel 2022]. We aim to mitigate inconsistencies in the shared embedding spaces of video and pose representations across different viewpoints. To achieve this, we introduce two geometric consistency regularizers, which are defined over each mini-batch as follows:

(1) *Cross-Modal Geometric Consistency Loss*: This loss minimizes discrepancies in similarity scores for video-pose pairs across different viewpoints. It is formulated as:

$$\mathcal{L}_{C-Geo} = \frac{1}{N}\sum_{n=1}^{N}\sum_{(p,q)\leq V}(\langle I_n^p, J_n^q\rangle - \langle J_n^p, I_n^q\rangle)^2, \tag{6.6}$$

where I_n^p and J_n^q represent video and pose embeddings, respectively, for viewpoint p and q. Also, $\langle I_n^p, J_n^q\rangle$ and $\langle J_n^p, I_n^q\rangle$ measure the similarity between video and pose embeddings for different viewpoints.

(2) *In-Modal Geometric Consistency Loss*: This loss ensures that similarity scores remain consistent across viewpoints across video and pose pairs. It is defined as:

$$\mathcal{L}_{I-Geo} = \frac{1}{N}\sum_{n=1}^{N}\sum_{(p,q)\leq V}(\langle I_n^p, I_n^q\rangle - \langle J_n^p, J_n^q\rangle)^2, \tag{6.7}$$

Here, $\langle I_n^p, I_n^q\rangle$ and $\langle J_n^p, J_n^q\rangle$ measure the similarity between video and pose pair embeddings, from two different viewpoints.

These constraints collectively enhance the geometric consistency of the learned embeddings across modalities and viewpoints.

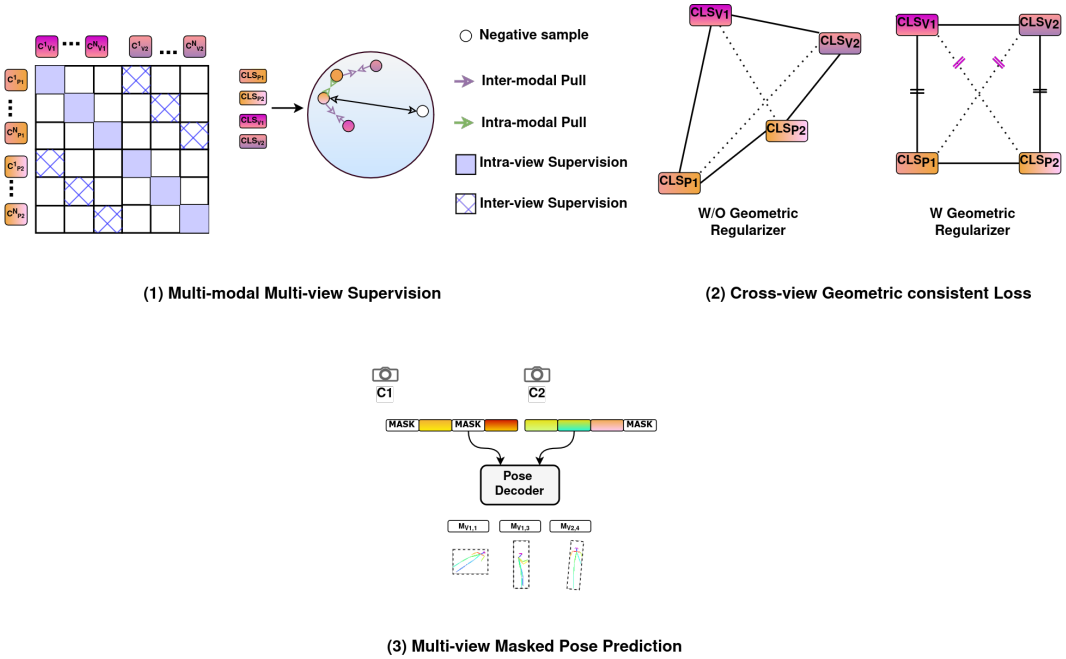

Figure 6.2: We use different pretraining objectives on the global representations of each modality and viewpoint.

Masked Pose Modeling

We follow the encoder-decoder design in MAE [He 2022a], where the transformer encoder focuses on representation learning while the decoder implements the reconstruction task. Our decoder takes as input the aligned pose features $\{\overline{\pi}^1_{1,1}, ..., \overline{\pi}^V_{N_p,T}\} \in \mathbb{R}^{V \times N_p \times T \times d}$ output by the encoder. The reconstruction target corresponds to the initial coordinates of randomly sampled pose tokens that have been masked before being fed to our encoder (see Fig. 6.2). We use the L_{Mask} loss as a simple mean-squared error loss between the predicted and target coordinates.

Finally, the total loss for all the pretraining objectives is defined as follows:

$$L_{Align} = L_{Con} + \lambda_1(L_{C-Geo} + L_{I-Geo}) + \lambda_2 L_{Mask}, \tag{6.8}$$

where λ_1 and λ_2 are hyperparameters controlling the importance of the geometric consistency and masked pose prediction regularizers.

6.2.4 Finetuning on Action Recognition

Our model is trained in two stages. Following the pretraining phase described earlier, we finetune the pretrained encoders for surgical action recognition. For each modality $M \in \{I, J\}$ and each available viewpoint $c \in [1, \ldots, C]$, we extract a global token CLS^c_M. We stack these tokens and perform an average pooling operation to obtain the overall global represen-

tation. This representation is then input into a two-layer multi-layer perceptron (MLP) to generate class probabilities. This adaptable representation enables us to utilize different viewpoints in our pretraining, finetuning, and testing framework.

6.3 Datasets

6.3.1 4D-OR Dataset

The 4D-OR dataset [Özsoy 2022] encompasses 10 simulated total knee replacement surgeries carried out in a medical simulation center under the supervision of orthopedic surgeons. The average recording duration is 11 minutes with 6,734 images per camera. The dataset is captured from 6 RGB-D Kinect cameras strategically mounted on the OR ceiling, ensuring complete coverage of the OR. Among the 6 available camera points of view, the sixth camera offers a different perspective. It is located on the ceiling, thus providing a quasi-bird's-eye view of the scene. The workflow in the dataset is a simulated and simplified version of an actual surgery, and the actors' roles are regularly rotated to introduce variability in the dataset. The cameras are fixed during all procedures, enabling the ablation experiments and, notably, the cross-view experiments in section 6.4.2.

6.3.2 Implementation details

We implement our method using PyTorch [Paszke 2019] based on the PySlowfast library. Our baseline model is the space-time MViT-S [Fan 2021] with MaskFeat pretraining [Wei 2022]. In all experiments, we use an input size of (8, 224, 224) and a token cube size of (2, 16, 16), which results in a total of 784 vision tokens. As mentioned in section 6.2.2, we set $N_p = 8$ as the maximum number of persons detected in a single frame for both datasets, resulting in a total of 64 pose tokens per viewpoint.

We exploit all available camera viewpoints in both datasets in our pretraining experiments. We use a stochastic gradient descent as our optimizer, with a cosine learning rate schedule over 50 epochs, with 0.01 as the initial learning rate and a weight decay of $1e^{-5}$. We use $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ for our pretraining experiments.

We finetune our models and baseline models using the AdamW optimizer [Loshchilov 2017] with a learning rate of $1e^{-4}$ and a batch size of 16 videos, applying cosine learning rate decay. In our experiments, we keep the weights of the first 12 layers of the video encoder frozen.

During training, images are resized to the shortest side of 256 pixels for pretraining and finetuning with a random crop to 224 pixels. For testing, images are resized to the shortest side of 224 pixels with a center crop.

Table 6.1: Accuracy (%) for surgical action recognition on the 4D-OR dataset using different models and pretraining strategies. Results are averaged over three seeds.

			Modalities		**# Cases**					
Model	**PT Method**	**PT Data**	**RGB**	**2D Pose**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
MViT-S	MaskFeat	K400	✓	✗	$50.8_{\pm 1.2}$	$63.6_{\pm 0.3}$	$72.3_{\pm 0.5}$	$76.2_{\pm 0.7}$	$79.8_{\pm 1.3}$	$84.6_{\pm 0.9}$
ViT-B	VMAE	K400	✓	✗	$48.8_{\pm 1.2}$	$62.7_{\pm 0.4}$	$71.5_{\pm 0.7}$	$74.0_{\pm 1.5}$	$77.3_{\pm 1.0}$	$81.3_{\pm 0.7}$
PCT-TF	None	N/A	✗	✓	$38.3_{\pm 2.4}$	$45.2_{\pm 1.1}$	$51.4_{\pm 0.7}$	$58.0_{\pm 1.2}$	$54.2_{\pm 1.1}$	$69.5_{\pm 0.8}$
PCT-MViT-S	MaskFeat	K400	✓	✓	$53.2_{\pm 1.5}$	$64.8_{\pm 0.8}$	$78.5_{\pm 2.4}$	$80.5_{\pm 2.1}$	$83.4_{\pm 1.2}$	$85.1_{\pm 0.7}$
MV-CLIP	CLIP*	4D-OR	✓	✓	$54.2_{\pm 1.6}$	$65.9_{\pm 1.5}$	$79.3_{\pm 1.8}$	$82.8_{\pm 1.4}$	$84.2_{\pm 1.3}$	$85.5_{\pm 0.7}$
PreViPS	**Ours**	**4D-OR**	✓	✓	$\mathbf{55.4}_{\pm 2.1}$	$\mathbf{68.2}_{\pm 1.5}$	$\mathbf{80.9}_{\pm 2.7}$	$\mathbf{85.2}_{\pm 1.4}$	$\mathbf{88.7}_{\pm 0.8}$	$\mathbf{89.6}_{\pm 0.4}$

Table 6.2: Accuracy (%) for surgical action recognition on the OR-AR dataset across different models and data fractions. Results are averaged over three seeds.

			Modalities		**Data %**				
Model	**PT Method**	**PT Data**	**ToF**	**2D Pose**	**5%**	**10%**	**20%**	**50%**	**100%**
MViT-S	MaskFeat	K400	✓	✗	$40.1_{\pm 1.4}$	$56.7_{\pm 0.3}$	$63.2_{\pm 0.5}$	$80.6_{\pm 0.7}$	$86.3_{\pm 1.3}$
ViT-B	VMAE	K400	✓	✗	$42.8_{\pm 0.8}$	$61.4_{\pm 0.4}$	$67.8_{\pm 0.7}$	$84.4_{\pm 1.5}$	$88.9_{\pm 1.0}$
PCT-TF	None	N/A	✗	✓	$38.5_{\pm 1.1}$	$40.7_{\pm 1.1}$	$51.4_{\pm 0.7}$	$58.0_{\pm 1.2}$	$64.2_{\pm 1.1}$
PCT-MViT-S	MaskFeat	K400	✓	✓	$45.2_{\pm 0.8}$	$65.6_{\pm 0.8}$	$70.0_{\pm 2.4}$	$84.4_{\pm 2.1}$	$89.4_{\pm 1.2}$
MV-CLIP	CLIP*	OR-AR	✓	✓	$54.5_{\pm 1.6}$	$66.4_{\pm 1.3}$	$71.8_{\pm 2.0}$	$85.2_{\pm 1.2}$	$90.7_{\pm 1.1}$
PreViPS	**Ours**	**OR-AR**	✓	✓	$\mathbf{58.4}_{\pm 1.3}$	$\mathbf{70.1}_{\pm 1.5}$	$\mathbf{73.3}_{\pm 2.7}$	$\mathbf{87.8}_{\pm 1.4}$	$\mathbf{92.3}_{\pm 0.8}$

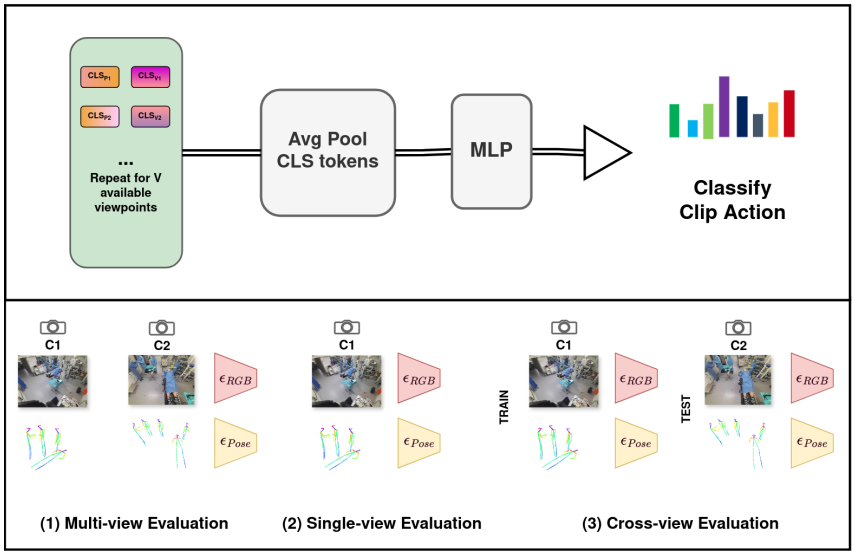

Figure 6.3: **Above:** We present our finetuning protocol, utilizing global representations from various modalities and viewpoints. **Below:** Additionally, we demonstrate the versatility of our approach, enabling us to train and test our methods using different viewpoints.

6.4 Experiments and Results

Baselines: In this section, we present a series of baselines that illustrate our pretraining strategy's advantages and pose-centric representations. We examine the performance of our method against five baselines:

- ◇ **MaskFeat** [Wei 2022] and **VideoMAE** [Tong 2022] are both state-of-the-art self-supervised approach used in general computer vision. These methods are *visual-only* approaches and do not use explicit pose information. The purpose of including these methods is to demonstrate the significant improvements that can be achieved by integrating pose information into our pretraining framework. This highlights the effectiveness of both our pose architecture and our unsupervised alignment objective in utilizing multi-modal information.

- ◇ **PCT-TF** (PCT + Transformer) method serves as a *pose-only* baseline to show the performance of our pose-based approach as it has not been explored yet for the task of surgical activity recognition from the external cameras.

- ◇ **PCT-MViT-S** is derived from a *dual-encoder architecture* without any pretraining. This architecture integrates our pose-based backbone (*PCT-TF*) with the video backbone (MViT-S). The video backbone has been pretrained using the Maskfeat method on the Kinetics-400

dataset. This combination demonstrates the advantages of incorporating appearance-based features alongside our pose representations, as pose information alone may not be sufficient to identify certain surgical actions.

- ◇ **MV-CLIP** utilizes the same architecture as PCT-MViT-S but further pretrains both encoders with a multi-view video-pose adaptation of the CLIP contrastive objective, referred to as CLIP* in section 6.2.3.

6.4.1 Data Efficient Transfer

Setup: We implement a data-efficient experimental protocol to demonstrate the advantages of our unsupervised training approach. We pretrain our *MV-CLIP* and *PreViPS* methods on each of the two datasets presented in sections 6.3.1 and 3.2, using all available viewpoints: 6 for 4D-OR and 4 for OR-AR, respectively. We finetune our pretrained model with progressively increasing amounts of labeled data. We use all available viewpoints for finetuning our models. We utilize the averaged global representations from every view and modality to create our clip representation.

To reduce potential bias in sampling videos from the dataset, we take each data sample three times and calculate the mean and standard deviation of the results. For both datasets, the testing and validation datasets remain unchanged. The data-efficient performance of the models is presented in Table 6.1 and Table 6.2. In Table 6.1, the "# Cases" column indicates the number of surgical cases used to fine-tune our models. These tables show that our pretraining significantly improves downstream task performance. Consistently, our PreViPS model outperforms the model trained from scratch across all label percentages on both datasets. Notably, as the amount of labeled data increases, the performance gap between the best video-based baseline and our method without pretraining narrows, demonstrating the effectiveness of our pretraining strategy.

6.4.2 Unimodal and Cross-View Evaluation

Our model adapts to various input modalities and viewpoints, allowing for unimodal and cross-view setups. The following section will explain our experimental setup for the cross-view, unimodal, and single-view experiments. We have conducted our experiments with 4D-OR because the camera setup is consistent, which helps us identify the cameras for our viewpoint ablation study.

Robustness to Viewpoint Shift

To assess the influence of varying camera viewpoints, we conduct three experiments. We test on one camera viewpoint in each experiment while using the other two for training. We focus on cameras 1, 4, and 6, as they provide a comprehensive overview of the scene with minimal overlap.

X-View Setup	$\{1,4\};6$	$\{1,6\};4$	$\{4,6\};1$
Accuracy Boost	**+6.4**	**+5.3**	**+3.5**

Table 6.3: Effectiveness of our alignment pretraining when holding out different viewpoints on 4D-OR. Performance increases are given in (%) for different Train-Test camera setups.

Scratch		**Pretrained**	
(P)	**(V)**	**(P)**	**(V)**
69.5 ± 1.5	85.1 ± 0.7	**71.9±1.3**	**86.8±0.6**

Table 6.4: Effectiveness of our alignment pretraining when finetuning on a single modality on 4D-OR. Top-1 Accuracy is given in (%) for both pose (P) and video (V) modalities.

As mentioned in section 6.3.1, presenting the 4D-OR dataset, one view (camera viewpoint 6) gives a very different perspective of the scene and has the most minor overlap with the rest. Our results (see Table 6.3) demonstrate that alignment pretraining significantly enhances performance across all cross-view setups. This improvement is particularly pronounced when testing on the top view from camera 6, demonstrating the effectiveness of our pretraining approach.

Unimodal Evaluation

In these experiments, we investigate whether our pretraining method enhances unimodal activity recognition performance for both the *pose-only* and *vision-only* single modality backbones. For the pose encoder, we initialize it using weights from our multi-modal pretrained network while excluding the weights of the video encoder. Similarly, for the video backbone, we disregard the pose encoder weights.

As shown in Table 6.4, our multi-view representation pretraining proves beneficial when finetuning on single modalities. When leveraging all available viewpoints, the pose backbone, which previously underperformed without pretraining, achieves a performance increase from 69.5% to 71.9%. Likewise, the video backbone's performance improves modestly from 85.1% to 86.8%. These results highlight the advantages of our pretraining approach for both modalities.

Single-view Evaluation

We conduct single-view experiments, where both training and testing occur from the same viewpoint. These experiments use the same viewpoints as those in the cross-view setup.

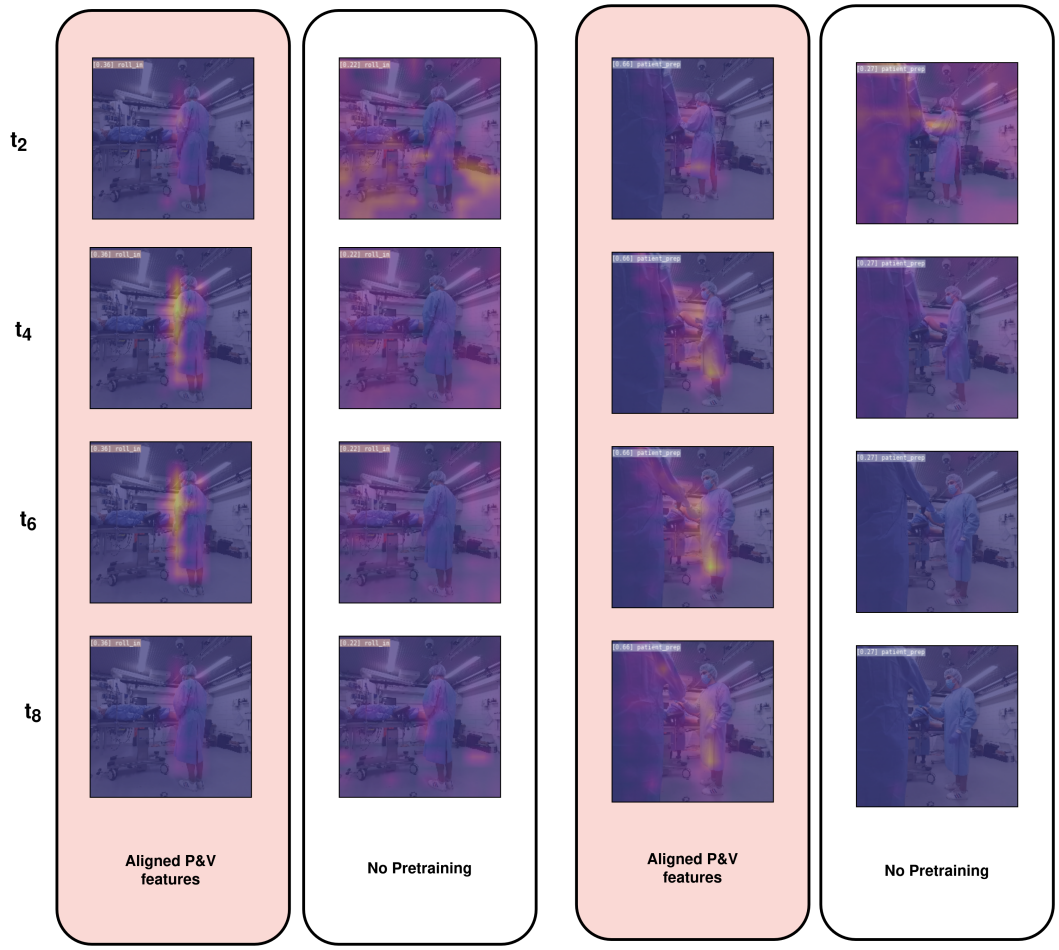

Figure 6.4: **GradCAM visualizations:** In the visualization of videos, brighter colors indicate higher attention. Notably, we observe that greater attention is assigned to moving body parts. The top row shows activation maps from our pretrained model with alignment objectives, while the bottom row displays results from the model trained without video-pose alignment.

Single View Setup	Camera 1	Camera 4	Camera 6
Accuracy Boost	**+5.6**	**+3.2**	**+5.1**

Table 6.5: Effectiveness of alignment pretraining when finetuning on a single view on 4D-OR. Performance increases are given in (%) is given in (%). Testing is done on the same camera viewpoint.

The results in Table 6.5 show a significant improvement in single-view control across different viewpoints when using our pretraining method. This highlights the benefits of multi-view representation learning, even when only a single camera is available for the downstream task.

6.4.3 Temporal Modeling

To enable in-context prediction and accurately model the sequential order of activities, we build on prior work [Sharghi 2020, Jamal 2022] and extend our finetuning approach with a recurrent neural network to capture global temporal information in the video. After extracting features from the training videos, each video is represented as $v_i = \{f_1, ..., f_T\}$, where f_T denotes global embeddings averaged across different views and modalities. These features are then processed using a Bidirectional Gated Recurrent Unit (BiGRU) [Chung 2014], producing an updated feature sequence $\overline{v_i} = \{\overline{f_1}, ..., \overline{f_T}\}$. The updated features are subsequently used for activity classification.

The results presented in Table 6.6 underscore the advantages of integrating enhanced temporal modeling. The asterisk (*) in Table 6.6 indicates the LABRADOR baseline [Özsoy 2023], which uses point cloud and scene graph information (using extra depth modality) and heuristic rules for activity prediction. A direct comparison is impossible since our method does not rely on semantic scene graph annotations or memory-heavy 3D point cloud data.

Accordingly, we consider LABRADOR to be an upper-bound baseline. Our video-pose-based approach attains competitive performance without requiring fine-grained scene graph supervision.

6.4.4 Ablation Study and Analysis

We perform extensive ablation experiments on the 4D-OR dataset to study the effect of our method's different contributions and design choices.

Effects of number of Views We perform an ablation study to validate the robustness of PreViPS to varying numbers of views during inference. We observe that the performance increases as more views are available for representation learning. A comparison is shown in Fig. 6.5 using only the pose encoder. This is intuitive as different views provide varying perspectives, which helps in recognizing actions better.

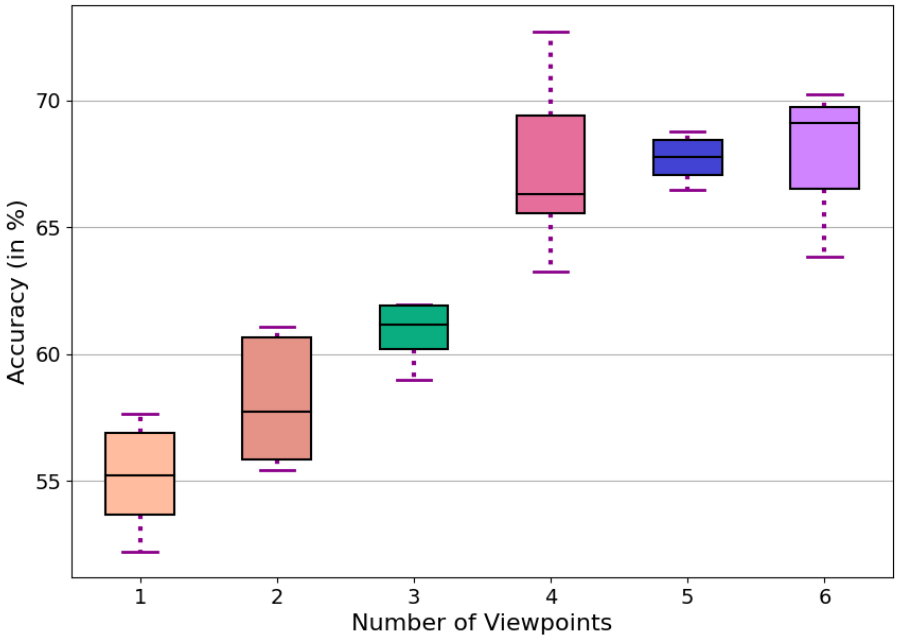

Figure 6.5: Box-plots showing Accuracy distributions from 4D-OR clip classification experiment for different camera viewpoints available. Ablation was run using only the pose modality as input.

Table 6.6: Results and comparison against baselines for OR surgical activity recognition on complete procedures. We provide the mAP, Precision, Accuracy, and F1 score.

Dataset	Model	mAP	Precision	Accuracy	F1
4D-OR	MaskFeat	84.2	82.7	86.0	80.7
	PCT-TF	77.1	75.3	78.3	75.6
	PCT-MViT-S	88.0	86.9	90.5	89.4
	MV-CLIP	90.8	89.3	92.0	90.6
	PreViPS	**92.9**	**91.7**	**94.2**	**93.4**
4D-OR	LABRADOR*	N/A	96.0	97.0	97.0
OR-AR	MaskFeat	89.5	87.3	89.8	88.6
	PCT-TF	75.2	73.6	77.0	75.6
	PCT-MViT-S	91.4	90.1	92.5	92.5
	MV-CLIP	92.4	90.7	93.5	93.1
	PreViPS	**93.6**	**92.0**	**95.4**	**94.3**

MV Contrastive	Geometric	Mask Pose	*Accuracy Drop*
✓	✓	✓	N/A
✓	✗	✗	**-4.1**
✓	✗	✓	**-3.3**
✓	✓	✗	**-1.8**

Table 6.7: Effect of keeping out different unsupervised objectives on PreViPS using 4D-OR. The multi-view contrastive objective is required in our ablation study.

Effect of pretraining objectives We conduct an ablation study to assess the contribution of each loss component in our pretraining objective, L_{Align}, as summarized in Table 6.7. The results show that all individual pretraining objectives are essential for optimal performance.

Component Ablation on Pose Token Representation. We analyze the components of our pose token representation, focusing on the choice of architecture for the pose tokenizer. Specifically, we compare a simple MLP-based tokenizer with the proposed PCT encoder. As shown in Table 6.8, the compositional representation of PCT leads to a significant performance improvement.

We integrate positional embeddings with pose embeddings to accurately encode detected human poses with their timestep, track ID, and viewpoint

Pose Token Ablation	***Accuracy* (%)**
MLP 2D Coords	60.3
PCT [Geng 2023]	**69.5**
W/O Pos Embed.	65.2
W Pos Embed.	**68.4**

Table 6.8: Effect of replacing the PCT [Geng 2023] pose tokenizer with a simple MLP baseline. Benefits from adding positional embeddings in our pose token representation.

ID. We conduct an ablation study on these positional encodings, and the results in Table 6.8 demonstrate the performance gains achieved by incorporating them into the pose representation.

Finally, Figure 6.4 visualizes the differences in activation maps between the pretrained model and a model trained without our alignment objectives. The pretrained model shows greater attention to moving body keypoints, indicating that our multi-modal pretraining approach effectively transfers 2D pose information to the vision encoder.

6.5 Conclusion

This chapter introduced *PreViPS*, a novel calibration-free pretraining framework for surgical activity recognition that integrates multi-view and multimodal signals. To our knowledge, *PreViPS* is the first approach to align 2D human pose and visual embeddings across uncalibrated camera views, a significant advancement given the practical constraints of operating rooms, where calibration is rarely feasible.

Our method enhances representation learning by combining discrete pose tokenization, geometric constraints across and within modalities, and masked pose modeling. These components collectively improve recognition performance in multi-view and single-view settings, while remaining lightweight and calibration-agnostic.

This work underscores the potential of self-supervised and geometry-aware learning in developing robust, scalable systems for surgical video understanding. PreViPS takes a meaningful step toward the practical deployment of intelligent systems in real-world surgical environments by eliminating the need for precise calibration or depth sensing.

Part II

Discussion and Conclusions

7

Conclusions and Future Work

> Choose a representation that can use unsupervised learning on unlabeled data, which is so much more plentiful than labeled data.
>
> Peter Norvig

Contents

This thesis presented a series of contributions to reduce supervision costs for surgical video understanding by leveraging multimodal self-supervised learning. We demonstrated how structured, semantically meaningful supervision can be derived from unannotated data by introducing abstract semantic modalities, such as object layouts, human pose, and spatial clustering derived from depth cues. These methods improve robustness and data efficiency in surgical scene analysis tasks, including activity recognition and semantic segmentation.

7.1 Summary of Contributions

Our contributions are unified under a common framework that uses cross-modal alignment and proxy objectives to extract rich visual representations with minimal annotation. Specifically, the proposed methodology can be parsed along three principal methodological axes, namely:

- **Geometric abstraction:** We introduced depth-based clustering as a structural proxy for unsupervised learning.
- **Object-centric modeling:** By leveraging object and clinician layouts, we proposed masking-based objectives to capture co-presence and spatial dependencies.
- **Pose-guided alignment:** Using off-the-shelf pose estimators, we aligned temporal segments across multiple views and modalities via contrastive pretraining strategies.

Together, these contributions move toward label-efficient learning in complex surgical environments, providing insights into how abstract information can structure visual representation learning.

7.1.1 Limitations

While this thesis proposes promising approaches for label-efficient surgical scene understanding, several limitations will need to be addressed in the future. These limitations point to areas where further research is necessary to ensure broader applicability, performance stability, and practical deployment in clinical environments.

- **Limited Generalizability Across Clinical Settings:** Although the dataset used in this work [Sharghi 2020] encompasses recordings from multiple operating rooms, all data were collected within a single hospital. Future studies could beneficially assess the proposed methods in a broader range of surgical environments, teams, and procedural standards to confirm and strengthen evidence for the robustness of abstract modality interactions in heterogeneous clinical contexts.

- **Performance Gaps in Abstract Supervision:** Techniques that rely on abstract representations, such as human pose or object layouts, still trail behind appearance-based models regarding raw predictive accuracy. This limitation stems less from architectural challenges and more from the lack of fine-grained information in abstract representations. Notably, appearance-based features derived from self-supervised frameworks like DINO [Caron 2021] have been shown to encode rich object-level semantics [Siméoni 2021, Wang 2023a, Wang 2023b, Arica 2024], suggesting an opportunity to combine these complementary sources of information, rather than viewing them as mutually exclusive, to enhance model performance further.

- **Increased Computational Overhead:** Using auxiliary modules such as object detectors and pose estimators introduces additional computational costs during training and inference. These components may hinder real-time deployment in ORs with limited computational resources. Although object-centric approaches that implicitly encode spatial structure (e.g., [Ding 2021]) offer potential solutions, striking the right balance between model efficiency and expressiveness remains an ongoing challenge.

7.2 Future Directions

Building upon the limitations discussed above, several promising research directions emerge that can strengthen the scalability, efficiency, and clinical relevance of multimodal learning systems in surgical environments.

7.2.1 Model Compression and Efficiency

As computational efficiency remains a key challenge for real-world deployment, future work should focus on model distillation and compression techniques. One promising approach is to distill knowledge from large-scale video encoders into smaller, more efficient abstract encoders. By compressing visual content into object-centric tokens [Qian 2024], these lightweight models can retain temporal reasoning capabilities while dramatically reducing computational costs. This is particularly beneficial in surgical settings, where hardware resources are limited and real-time inference is critical.

7.2.2 End-to-End Learning of Semantic Modalities

Our current pipelines rely on external modules (e.g., object detectors, pose estimators) to extract abstract modalities, which introduces latency and system complexity. A key future direction involves developing end-to-end architectures that predict abstract semantic representations (e.g., semantic maps, object layouts, pose graphs) parallel to visual features. This integrated design would improve efficiency, facilitate deployment, and support

joint optimization of semantic and visual representations.

7.2.3 Multimodal Expansion and Cognitive Integration

Extending the modality space beyond vision and geometry offers another fruitful avenue. Signals like audio cues, eye-tracking data, and haptic or tactile feedback can enhance task definition and situational awareness. Integrating these modalities could lead to a more robust understanding of cognitive and procedural contexts. Furthermore, interdisciplinary collaboration with cognitive scientists, human factors experts, and psychologists may guide the design of semantically grounded representations, especially for modeling intention, interaction, or mental workload in high-stakes environments.

7.2.4 Bridging Vision and Language for Cross-Modal Understanding

The discrete and semantically rich nature of abstract modalities (e.g., pose, object-centric layouts) positions them as ideal intermediaries between visual and linguistic representations. This enables opportunities for vision-language pretraining tailored to the surgical domain, where models learn to align visual scenes with textual instructions, narration, or commands. Such capabilities could unlock applications in robotic scene comprehension and autonomous assistance. For example, object-aware models could be used for vision-language navigation tasks, such as facilitating automatic surgical robot docking, a foundational step toward Level 2 autonomy, where robots execute predefined setup tasks in response to operator input [Yang 2017]. This aligns with recent advancements in cross-modal learning [Zhang 2024] and may serve as a bridge between static perception and interactive AI systems.

7.2.5 Human–Robot Interaction and Clinical Translation

Real-world improvements depend heavily on effective human–robot collaboration. Recent breakthroughs, such as autonomous robots performing suturing nearly as well as human experts [Rivero-Moreno 2024], and robotic surgical assistants capable of anticipating surgeons' instrument needs [Wagner 2024, Li 2024], highlight how abstract semantic concepts can be practically integrated into robotic systems. Moving forward, research should focus on multimodal learning within flexible autonomy frameworks to enhance robots' ability to transparently understand human intent, offer predictive support, and share tasks ergonomically. By combining sophisticated semantic understanding with intuitive interfaces, surgical robots can transform from passive tools into proactive partners, significantly improving safety, efficiency, and the overall experience for surgeons.

7.2.6 Summary of Key Opportunities

Based on the discussion above, several opportunity areas are particularly compelling:

- **Reducing computational complexity:** Through model distillation and object-token compression [Qian 2024], future models can become significantly more efficient without sacrificing temporal understanding, crucial for clinical adoption.
- **Enabling deployment through architectural simplification:** End-to-end models that natively output semantic representations can eliminate dependencies on sequential external modules, improving inference time and reducing integration effort.
- **Enhancing semantic understanding across modalities:** Object-centric abstraction is a natural bridge between vision and language. This opens the door for multimodal applications in surgical robotics, real-time decision support, and task-guided automation.

7.3 Conclusion

This thesis has demonstrated that abstract modalities combined with visual data in a self-supervised setting offer a promising path toward efficient, scalable surgical scene understanding. By grounding models in semantic priors and structural cues, we move closer to the practical deployment of AI systems in complex, real-world OR environments. Future work must focus on generalization, computational efficiency, and ergonomically sound integration into clinical workflows.

A

Résumé en français

A.1 Introduction

L'irruption de l'Intelligence Artificielle (IA) dans de nombreux secteurs industriels s'inscrit dans une quatrième révolution industrielle, nourrie par la disponibilité croissante de données massives et hétérogènes. Le domaine médical, et plus particulièrement le bloc opératoire, bénéficie de cette dynamique avec l'intégration de capteurs visuels et de systèmes robotiques.

Cependant, les signaux visuels complexes issus de ces environnements ne sont pas directement exploitables par les modèles standards d'apprentissage profond, souvent préentraînés sur des bases de données très éloignées du contexte chirurgical. La compréhension automatique des vidéos de chirurgie ouvre pourtant des perspectives majeures : amélioration de l'ergonomie, standardisation des protocoles de communication et réduction de la charge cognitive des cliniciens.

Dans ce contexte, nous explorons l'utilisation de représentations dites *abstraites* [Liang 2022] (ex. pose humaine, superpixels, boîtes englobantes d'objets). Ces modalités, dérivées mais plus structurées que les signaux bruts, permettent de mieux généraliser tout en réduisant le besoin en annotations expertes.

A.2 Problématique et objectifs

Les approches actuelles d'analyse du workflow chirurgical reposent sur des réseaux profonds supervisés, très dépendants de larges bases de données annotées. Or, dans le contexte du bloc opératoire :

- l'annotation nécessite une expertise médicale coûteuse et rare ;
- la préservation de la vie privée limite la collecte et le partage des données ;
- la variabilité des configurations (procédures, hôpitaux, caméras) complique la généralisation.

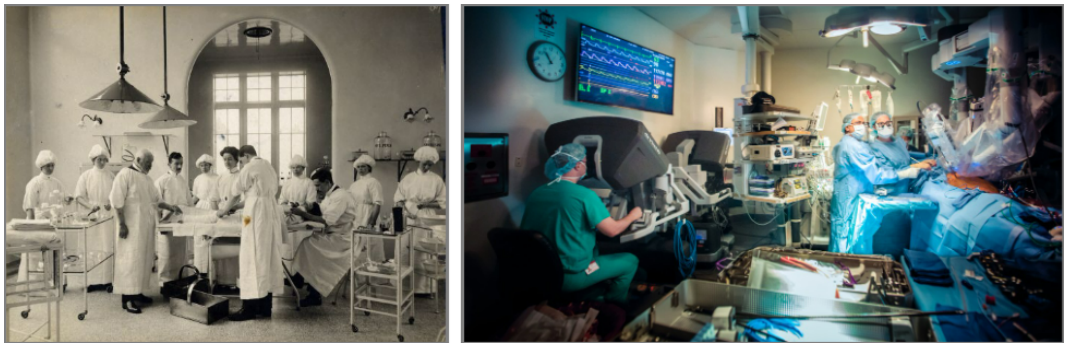

Figure A.1: Cette figure illustre l'évolution des pratiques chirurgicales du début du XX^{e} siècle à aujourd'hui. Elle met également en évidence les transformations des salles d'opération, marquées par la transition vers un environnement plus moderne, hautement technologique et encombré. D'après [Lefkowitz 2018].

L'objectif de cette thèse est de développer des méthodes d'**apprentissage multimodal auto-supervisé** exploitant à la fois les signaux visuels bruts (RGB, profondeur) et les modalités abstraites (pose, objets, superpixels). Ces représentations visent à :

(i) réduire la dépendance aux annotations manuelles (*label-efficiency*) ;

(ii) améliorer la robustesse aux changements de points de vue et de domaines ;

(iii) préserver la confidentialité des patients et cliniciens.

A.3 Contributions principales

A.3.1 Analyse de jeux de données multimodaux

Cette thèse s'appuie sur deux jeux de données complémentaires dédiés au **suivi du workflow au bloc opératoire**: **OR-AR** [Sharghi 2020] pour la reconnaissance d'activités à l'échelle de la salle, et **OR-Seg** [Li 2020a] pour la segmentation pixel à pixel des composants robotiques. Contrairement aux benchmarks endoscopiques, ces ressources couvrent la *vue salle*, des *modalités compatibles vie privée* (ToF/IR) et des *annotations adaptées* aux tâches de contexte.

OR-AR Capteurs *Time-of-Flight* (profondeur + intensité IR) montés sur quatre chariots (deux caméras/chariot, base $\approx$70 cm) positionnés pour limiter les occultations. Le corpus regroupe **400** vidéos issues de **103** procédures, **27** chirurgiens, **30** types d'actes, sur **2** salles pendant **2** ans. Dix activités structurent le déroulé opératoire (préparation stérile,

Figure A.2: Visualisation des dix activités de suivi du flux de travail annotées dans OR-AR [Sharghi 2020], avec leurs occurrences respectives. D'après [Sharghi 2020].

roll-in/out patient, préparation patient, roll-up/out robot, docking/undocking, chirurgie robotique, fermeture). Suivant les protocoles de la littérature [He 2022b, Jamal 2023b], la *préparation stérile* est souvent exclue des évaluations (*class imbalance*/ambiguïté). Un sous-ensemble, **OR-Det**, annote **19k** images (20 vidéos) en boîtes *objets/personnes* : brancard, tables stérile/non-stérile, PSC, VSC, table d'opération et cliniciens support clé pour le raisonnement *"object-centric"*.

OR-Seg Données ToF collectées en environnement de développement clinique, avec scénarios de laparoscopie robot-assistée. Deux volets: **mono-vue** (7 980 images) et **multi-vues** (capteurs fixés sur le PSC). Les classes couvrent les principaux sous-composants du système *da Vinci*; le fort *déséquilibre* de pixels rend la **fwIoU** informative en plus de la **mIoU**.

A.4 Auto-supervision sur cartes de profondeur par superpixels

Nous présentons un **cadre d'auto-supervision** destiné à l'analyse du contexte au bloc opératoire, reposant uniquement sur des cartes de profondeur issues de capteurs *temps de vol* (ToF). L'objectif est double : **réduire la dépendance aux annotations expertes** et **préserver la confidentialité** (pas d'images RGB de visages), tout en améliorant la compréhension de la scène pour la segmentation sémantique et la reconnaissance d'activités.

Idée directrice. Les **distances géométriques** entre entités de la salle (bras robotisés, table, équipe) demeurent stables quel que soit le point de vue.

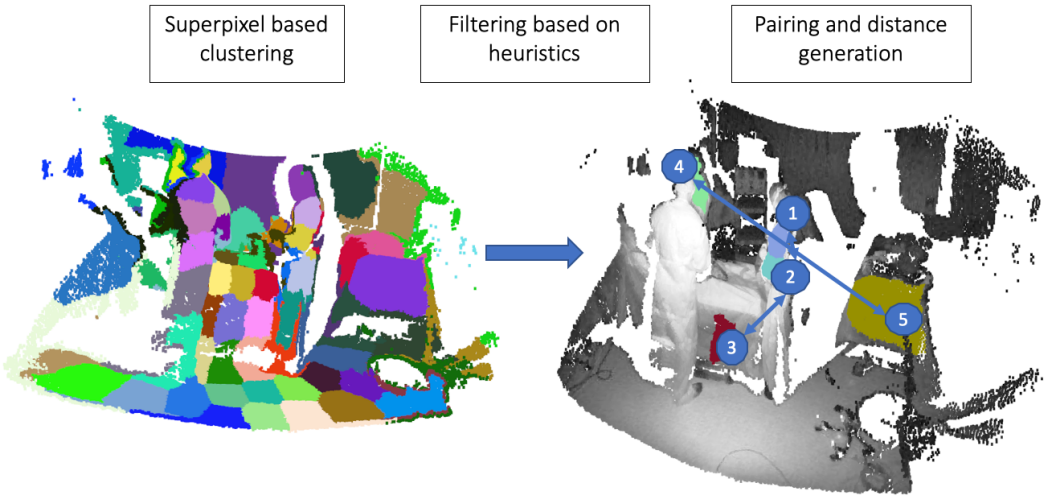

Figure A.3: Pretext task annotation generation process using SLIC [Achanta 2012] superpixel segmentation.

Nous exploitons cette invariance en formulant une **tâche prétexte** qui apprend à *prédire la distance euclidienne 3D* entre *superpixels* homogènes extraits des cartes de profondeur.

Chaîne de traitement

1. **Sur-segmentation par superpixels** (SLIC) sur l'image de profondeur lissée ; sélection de régions compactes et peu bruitées (critères de convexité, faible variance de profondeur, peu de valeurs manquantes).

2. **Projection 3D** des superpixels via les paramètres intrinsèques caméra ; calcul de la **distance entre centroïdes** pour générer des *pseudo-étiquettes* de distances.

3. **Apprentissage auto-supervisé** : un extracteur de caractéristiques (type ResNet) suivi d'un simple *décodeur* apprend des représentations dont la distance L_2 en espace des caractéristiques *reproduit* la distance 3D entre superpixels ($L_{\text{prétexte}} = |\|h_1 - h_2\|_2 - d_{3D}|$).

4. **Finetuning** sur deux tâches cibles : *segmentation sémantique* (mIoU, fwIoU) et *reconnaissance d'activités* (mAP), avec différents taux d'annotation (2 % à 100 %).

Jeux de données et tâches Évaluation sur deux ressources publiques de vision vue salle : **OR-Seg** (segmentation des composants du robot *da Vinci*) et **OR-AR** (activités au niveau de la salle). Les métriques suivent les protocoles d'origine : mIoU/fwIoU pour la segmentation, mAP pour l'activité.

Références de comparaison Nous comparons à des méthodes d'auto-supervision reconnues **adaptées aux cartes de profondeur** : prédiction d'angles de rotation (RotNet) et contraste prédictif (CPC v2), ainsi qu'à un apprentissage **depuis zéro** (sans pré-apprentissage). Les outils et hyper-paramètres sont harmonisés (durées d'entraînement identiques, sélection sur validation).

Résultats principaux

- **Gain net en faible supervision** : notre prétexte surpasse l'entraînement depuis zéro et devance RotNet/CPC v2 lorsque la fraction annotée est faible (2–20 %), *pour la segmentation et l'activité*, et ce *quelle que soit* la taille de l'extracteur (type ResNet-18/50).
- **Convergence des performances** lorsque la quantité d'annotations augmente : l'écart se réduit et devient comparable aux autres auto-supervisions à 50–100 %.
- **Efficacité d'annotation** : en reconnaissance d'activités, une performance mAP équivalente à un modèle entraîné à 50 % d'annotations peut être atteinte avec **environ la moitié** des annotations (ex. 5 % vs 50 % dans nos essais).
- **Significativité statistique** : tests de Wilcoxon avec corrections (Dunnett, Bonferroni-Holm) indiquent des améliorations significatives ($p \ll 0.05$) dans les régimes peu annotés (jusqu'à 20 % pour la segmentation, 10 % pour l'activité), et même $p < 0.01$ sur les plus faibles fractions en segmentation.

Analyse qualitative Les projections t-SNE des caractéristiques de superpixels montrent des **amas cohérents par entité** (ex. table, éclairage plafonnier, silhouettes humaines) *sans supervision*, signe que la tâche prétexte capte une structuration géométrique utile.

Forces et limites Forces : (i) **confidentialité** respectée (profondeur seule), (ii) **invariance au point de vue** via la géométrie, (iii) **frugalité en annotations** et transfert aux deux tâches. Limites : résultats en segmentation encore en deçà des meilleures architectures spécialisées, car nous privilégions des **architectures volontairement simples** pour isoler l'effet du prétexte. Un couplage avec des têtes de segmentation plus avancées devrait **améliorer le plafond de performance**.

Conclusion. L'**apprentissage auto-supervisé guidé par la géométrie** sur cartes de profondeur, via distances entre superpixels, constitue une voie efficace pour doter le bloc opératoire d'une **compréhension de contexte** économe en annotations et compatible avec la confidentialité. Cette brique servira, dans les chapitres suivants, de **socle** à des approches plus riches

intégrant *modalités abstraites* (objets, pose) et *multi-vues* pour la reconnaissance d'activités.

A.5 Reconnaissance d'activités centrée-objets avec pré-apprentissage masqué

Nous étudions une **approche centrée-objets** pour reconnaître les activités au bloc opératoire à partir de vidéos basse résolution (RGB ou ToF). L'hypothèse clé est que la **disposition géométrique** des dispositifs (PSC, VSC, table d'opération, brancard, tables stérile/non stérile) et des **cliniciens** porte une information discriminante forte. Deux étapes complémentaires sont proposées.

Étape 1 — ST(OR)^2 : raisonnement objet–temps par réseaux simples. Nous détectons personnes et objets à l'aide de détecteurs spécialisés entraînés sur *OR-Det* (sous-ensemble annoté d'OR-AR). Pour chaque extrait court, chaque détection est décrite par:

- une **représentation spatiale** (centre, largeur, hauteur de la boîte),
- une **représentation sémantique** (catégorie).

Ces deux informations sont projetées par des MLP, puis **agrégées par catégorie** afin d'éviter le suivi instance-par-instance. Un module temporel compact raisonne ensuite *au fil des images*, puis *entre catégories*, pour produire un vecteur d'extrait utilisé en classification d'action. Cette représentation objet peut être **fusionnée a posteriori** avec des descripteurs d'apparence globaux issus d'un réseau vidéo (p. ex. I3D [Carreira 2017]). Sur la *reconnaissance d'activités* (segmentation longue), l'ajout de ST(OR)^2 aux descripteurs d'apparence améliore la mAP sur OR-AR.

Limite. L'agrégation par catégorie supprime l'information des **multiples instances humaines**, et l'ordre des objets dans une image ne devrait pas influencer la prédiction : il faut un modèle **invariant par permutation**.

Étape 2 — ORDynaRe : transformeur spatio-temporel et pré-apprentissage masqué "centré-objets" Nous introduisons **ORDynaRe**, qui traite une *séquence de tokens objets* par un **transformeur** spatio-temporel, ce qui apporte :

1. **Invariance par permutation** des objets au sein d'une image,
2. **Gestion naturelle des multi-instances** (notamment les cliniciens),
3. **Pré-apprentissage auto-supervisé** adapté aux objets

Chaque token combine: (i) position (boîte), (ii) catégorie, (iii) caractéristiques de région issues du détecteur; un **codage temporel** est ajouté.

Le **pré-apprentissage masqué** consiste à *cacher* une piste courte d'un objet (toutes les images intermédiaires) en ne gardant que le premier et le dernier token comme contexte, puis à **prédire** pour les tokens masqués *à la fois* la **catégorie** et la **position** (et, en régularisation, les caractéristiques initiales). La perte combine entropie croisée (catégorie) et régression de boîtes (L1 + gIoU), sur les seuls tokens masqués.

Intégration et affinage Après pré-apprentissage, on affine le transformeur pour la **classification d'actions** d'extraits (token spécial [CLS]), puis pour la **reconnaissance d'activités** (segmentation longue) via un décodeur temporel (p. ex. GRU). Les vecteurs centrés-objets peuvent être **combinés** aux descripteurs d'apparence pour accroître la robustesse.

Jeu de données et mesures Les expériences sont menées sur **OR-AR** [Sharghi 2020] (9 activités retenues, fort déséquilibre de durées), avec détecteurs entraînés sur **OR-Det**. Les mesures suivent les usages: *exactitude top-1* pour la classification d'actions (extraits), *mAP* pour la reconnaissance d'activités (vidéos longues).

Résultats

- **Apport du pré-apprentissage** : ORDynaRe pré-entraîné surpasse la version sans pré-apprentissage *à toutes les fractions annotées*, avec des gains marqués en faible supervision (5–20 %).
- **Complémentarité apparence/objets** : la fusion avec des descripteurs globaux (I3D) améliore encore la classification d'actions et la segmentation longue, et peut dépasser les méthodes purement globales lorsque l'annotation est limitée.
- **Interprétabilité** : les cartes d'attention du token [CLS] mettent en évidence les **objets déterminants** (p. ex. PSC pour *roll-up/roll-back*).

Conclusion. En structurant la vidéo par **objets** et en apprenant à **reconstituer** leurs catégories et positions lorsqu'ils sont masqués, nous obtenons des représentations temporelles efficaces, **économes en étiquettes** et **complémentaires** aux indices d'apparence. Cette brique centrée-objets prépare l'alignement avec d'autres modalités abstraites (p. ex. la pose) et s'intègre aux approches multi-vues présentées ensuite.

A.6 Alignement vidéo–pose multivues sans calibration

Nous présentons **PreViPS**, un cadre de **pré-apprentissage multimodal multivues** *sans calibration* pour la reconnaissance d'activités au bloc opératoire. L'idée centrale est d'**aligner finement la pose humaine** et les **indices visuels** issus de caméras non calibrées, afin d'exploiter les gestes des cliniciens tout en restant robuste aux changements de point de vue.

Motivation. Les approches courantes de reconnaissance d'activités (découpage en extraits puis modélisation temporelle) utilisent surtout des indices globaux d'apparence et ignorent souvent les **mouvements fins** des cliniciens ou exigent des **montages calibrés** et des nuages de points coûteux. Or, la pose 2D *vue salle* est aujourd'hui fiable et porteuse d'une sémantique gestuelle déterminante.

Principe architectural. PreViPS s'appuie sur un **double encodeur** (vidéo et pose) :

- **Branche vidéo** : encodeur vidéo de type ViT appris par masquage (MaskFeat), produisant un token global [CLS] par vue.
- **Branche pose** : les poses 2D (17 points) détectées par vue sont **discrétisées en tokens compositionnels** (PCT), ce qui transforme des coordonnées continues en *tokens* robustes aux occultations ; des **codages positionnels** (temps, identité de suivi, identifiant de vue) structurent la séquence. Un transformeur produit un [CLS] par vue.

Objectifs de pré-apprentissage. Nous alignons les représentations vidéo–pose et favorisons l'invariance au point de vue par trois familles de contraintes :

1. **Contraste multimodal multivues** (*type CLIP adapté*) : rapprocher, pour un même instant, la vidéo d'une vue et la pose de la même scène (toutes vues), tout en éloignant les paires négatives ; déclinaisons *inter-modalité* (vidéo–pose) et *intra-modalité* (vidéo–vidéo, pose–pose).
2. **Cohérence géométrique** : pénaliser les incohérences de similarité *entre modalités* et *au sein d'une modalité* selon les vues, pour stabiliser l'espace partagé.
3. **Masquage de tokens de pose** : masquer une partie des tokens de pose et **reconstruire** les coordonnées manquantes via un petit décodeur, afin de renforcer la représentation structurale des gestes.

La perte totale combine ces trois contributions avec des pondérations simples.

Affinage pour la tâche aval. Après pré-apprentissage, on **affine** les encodeurs pour la *classification d'actions* (extraits multivues) puis pour la *reconnaissance d'activités* (vidéos longues) en agrégeant les [CLS] de toutes les vues et, si besoin, avec un **décodeur temporel** (BiGRU).

Jeux de données et protocole. Évaluations sur **4D-OR** (6 caméras plafonnières, simulation guidée) et **OR-AR** (4 vues ToF en salle réelle). Nous étudions : (i) **parcimonie d'annotations** (5–100 %), (ii) **robustesse inter-vues** (apprentissage sur certaines caméras, test sur une autre), (iii) **monomodale** (vidéo seule ou pose seule), (iv) **monovue** (apprentissage et test sur la même vue).

Résultats (synthèse).

- **Données rares** : PreViPS dépasse les pré-apprentissages visuels seuls (MaskFeat, VideoMAE) et les variantes sans alignement pose–vidéo, avec des gains nets de 5–20 % d'annotations.
- **Changement de vue** : l'alignement multivues *sans calibration* améliore sensiblement les performances en **cross-view** (jusqu'à +6 pts selon la vue tenue à l'écart), y compris sur la vue zénithale la plus différente.
- **Monomodale** : le pré-apprentissage multimodal *bénéficie* aussi aux réglages **pose seule** et **vidéo seule** (+2 à +1.5 pts env.), montrant un **transfert croisé** utile.
- **Monovue et temporel** : en monovue, les gains restent marqués (+3 à +6 pts) ; l'ajout d'un décodeur temporel (BiGRU) consolide la chronologie des phases et améliore toutes les mesures (mAP, exactitude, F1).

Conclusion. PreViPS montre qu'un **alignement vidéo–pose multivues**, combinant **tokens de pose** et **contraintes géométriques**, apporte des représentations **robustes** et **parcimonieuses en étiquettes** pour la reconnaissance d'activités en salle d'opération, *sans* calibration ni 3D lourde. Ce cadre se prête naturellement aux extensions **multi-modalités abstraites** (pose + objets) et aux déploiements réalistes à **caméra unique**.

A.7 Conclusion

A.7.1 Synthèse et limites

Cette thèse a introduit des méthodes d'auto-supervision pour la compréhension de vidéos chirurgicales en exploitant des modalités abstraites telles que le regroupement spatial basé sur la profondeur, les dispositions d'objets et la pose humaine. Ces modalités offrent une supervision sémantiquement riche sans annotations manuelles, améliorant la robustesse et l'efficacité en données pour des tâches comme la reconnaissance d'activités et la segmentation sémantique. Nos contributions se structurent autour de trois axes principaux : (i) **l'abstraction géométrique** grâce au regroupement non supervisé de cartes de profondeur, (ii) **le raisonnement centré-objets** via des objectifs de masquage et de co-présence, et (iii) **l'alignement guidé par la pose** permettant de contraindre des représentations issues de caméras multivues non calibrées. Ensemble, ces travaux montrent comment l'information structurelle peut guider efficacement l'apprentissage multimodal de représentations.

Plusieurs limites demeurent néanmoins. D'une part, la **généralisation** à travers des hôpitaux et pratiques chirurgicales hétérogènes reste à démontrer. D'autre part, les modalités abstraites, bien que sobres en annotations, restent moins performantes que les caractéristiques d'apparence

pour la précision brute, ce qui incite à envisager des approches hybrides combinant les deux sources. Enfin, la dépendance à des modules externes (détecteurs d'objets, estimateurs de pose) introduit des **coûts computationnels** qui limitent l'utilisation en temps réel. Trouver le bon équilibre entre efficacité et expressivité constitue donc un défi majeur.

A.7.2 Perspectives

À partir de ces constats, plusieurs pistes de recherche apparaissent prometteuses. Une première priorité concerne **l'efficacité computationnelle** : la distillation de connaissances et la compression de grands encodeurs vers des modèles légers basés sur des jetons objets pourraient permettre une utilisation en temps réel au bloc opératoire. Une deuxième voie est le développement d'architectures **de bout-en-bout**, capables de prédire directement des abstractions sémantiques (cartes, graphes de pose, dispositions d'objets), sans dépendre de modules séquentiels, ce qui simplifierait le déploiement. Une troisième perspective est l'**extension multimodale**, en intégrant des signaux comme l'audio, le regard ou les retours haptiques, afin d'améliorer la compréhension contextuelle et la modélisation cognitive.

Une opportunité majeure réside également dans le **pont entre vision et langage** : la nature discrète et symbolique des modalités abstraites en fait des médiateurs naturels entre vidéo et instructions textuelles. Cela ouvre la voie à des pré-apprentissages vision–langage adaptés au domaine chirurgical, avec des applications en assistance robotique (navigation, positionnement automatique) ou en interfaces homme–machine plus intuitives. Enfin, les progrès récents en autonomie robotique montrent déjà le potentiel d'un soutien prédictif et proactif. Intégrer des représentations sémantiques riches à ces systèmes pourrait transformer les robots chirurgicaux en partenaires ergonomiques, améliorant à la fois l'efficacité, la sécurité et l'expérience des équipes au bloc.

En résumé, cette thèse montre que les modalités abstraites, combinées à l'auto-supervision visuelle, constituent une voie prometteuse vers une compréhension de scène chirurgicale plus économe en annotations et plus robuste. Les prochaines étapes devront viser la généralisation inter-établissements, la réduction des coûts computationnels et l'intégration dans des systèmes réellement utilisables en contexte clinique.

Acronyms

- **AI**: **A**rtificial **I**ntelligence
- **AP**: **A**verage **P**recision
- **BERT**: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- **CAI**: **C**omputer **A**ssisted **I**nterventions
- **CAS**: **C**ontext **A**ware **S**ystems
- **CNN**: **C**onvolutional **N**eural **N**etwork
- **CRF**: **C**onditional **R**andom **F**ield
- **DTW**: **D**ynamic **T**ime **W**arping
- **fps**: **f**rames **p**er **s**econds
- **GRU**: **G**ated **R**ecurrent **U**nit
- **GPU**: **G**raphics **P**rocessing **U**nit
- **HMM**: **H**idden **M**arkov **M**odel
- **HOG**: **H**istogram of **O**riented **G**radients
- **HPE**: **H**uman **P**ose **E**stimation
- **IoU**: **I**ntersection **o**ver **U**nion
- **LSTM**: **L**ong **S**hort-**T**erm **M**emory
- **MAE**: **M**asked **A**uto **E**ncoder
- **MIS**: **M**inimally **I**nvasive **S**urgery
- **MLP**: **M**ulti **L**ayer **P**erceptron
- **OR**: **O**perating **R**oom
- **PSC**: **P**atient **S**ide **C**art
- **REST**: **Re**presentational **S**tate **T**ransfer
- **RGBD**: **R**ed **G**reen **B**lue + **D**epth
- **SAR**: **S**ynthetic **A**perture **R**adar
- **SDS**: **S**urgical **D**ata **S**cience

- **SSL**: **S**elf **S**upervised **L**earning
- **SVM**: **S**upport **V**ector **M**achine
- **TCN**: **T**emporal **C**onvolutional **N**etwork
- **ToF**: **T**ime **o**f **F**light
- **ViT**: **Vi**sion **T**ransformer
- **VSC**: **V**ision **S**ide **C**art

B

Appendices

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

Dissecting Self-Supervised Learning Methods for Surgical Computer Vision

Sanat Ramesh[a,c,1], Vinkle Srivastav[a,1,*], Deepak Alapatt[a,1], Tong Yu[a,1], Aditya Murali[a], Luca Sestini[a,d], Chinedu Innocent Nwoye[a], Idris Hamoud[a], Saurav Sharma[a], Antoine Fleurentin[b], Georgios Exarchakis[a,b], Alexandros Karargyris[a,b], Nicolas Padoy[a,b]

[a]*ICube, University of Strasbourg, CNRS, Strasbourg 67000, France*
[b]*IHU Strasbourg, Strasbourg 67000, France*
[c]*Altair Robotics Lab, Department of Computer Science, University of Verona, Verona 37134, Italy*
[d]*Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano 20133, Italy*

The field of surgical computer vision has undergone considerable breakthroughs in recent years with the rising popularity of deep neural network-based methods. However, standard fully-supervised approaches for training such models require vast amounts of annotated data, imposing a prohibitively high cost; especially in the clinical domain. Self-Supervised Learning (SSL) methods, which have begun to gain traction in the general computer vision community, represent a potential solution to these annotation costs, allowing to learn useful representations from only unlabeled data. Still, the effectiveness of SSL methods in more complex and impactful domains, such as medicine and surgery, remains limited and unexplored. In this work, we address this critical need by investigating four state-of-the-art SSL methods (MoCo v2, SimCLR, DINO, SwAV) in the context of surgical computer vision. We present an extensive analysis of the performance of these methods on the Cholec80 dataset for two fundamental and popular tasks in surgical context understanding, phase recognition and tool presence detection. We examine their parameterization, then their behavior with respect to training data quantities in semi-supervised settings. Correct transfer of these methods to surgery, as described and conducted in this work, leads to substantial performance gains over generic uses of SSL - up to 7.4% on phase recognition and 20% on tool presence detection - as well as state-of-the-art semi-supervised phase recognition approaches by up to 14%. Further results obtained on a highly diverse selection of surgical datasets exhibit strong generalization properties. The code is available at https://github.com/CAMMA-public/SelfSupSurg.

Keywords: Self-supervised learning; Semi-supervised learning; Surgical computer vision; Deep learning; Endoscopic videos; Laparoscopic cholecystectomy

1. Introduction

Automatic analysis and interpretation of visual signals from the operating room (OR) is the primary concern of surgical computer vision, a fast-growing discipline that is expected to play a major role in the development of reliable decision support systems for surgeons (Maier-Hein et al., 2017). Recent developments in the field have indeed resulted in increasingly refined vision algorithms; however, a majority of these studies have only been conducted on datasets containing small amounts of recorded procedures, all of which have been manually annotated by clinical experts. In future developments, much larger quantities of data will be required in order to account for variations in anatomy, patient demographics, clinical workflow, surgical skills, instrumentation, and image acquisition (Maier-Hein et al., 2022).

For that purpose, raw video data can be supplied on a very large scale by laparoscopic surgeries, since they are guided by intra-abdominal video streams: in the United States, nearly 1M laparoscopic cholecystectomies are performed each year, resulting in approximately 630k hours of footage for just this one type of procedure. Yet, datasets used for training current surgical vision models remain disproportionately small. For example, Cholec80 (Twinanda et al., 2016b), one of the most popular datasets in the field (Maier-Hein et al., 2017), hardly exceeds 50 hours of recordings. Apart from medico-legal constraints, the critical factor leading to this sparsity of data is the reliance on manual annotations. While labels for natural images can be easily supplied by the general public, surgical annotations usually require clinical expertise. As a result, the fully supervised approach - i.e. training models with entirely annotated datasets - may prove to be unsustainable in surgical computer vision.

In computer vision, an alternative has emerged in the form of Self-Supervised Learning (SSL) (Jing and Tian, 2021). Considerable progress has been made in this area, with increasingly refined methods for extracting rich vector representations from images without labels, using only the raw pixel data. This research topic has so far not been thoroughly explored in surgical applications. In the few self-supervised training tasks proposed by the community, learning from the visual content itself is generally de-emphasized in favor of utilizing other available sources of information - for example time (Funke et al., 2018; Yengera et al., 2018), stereoscopy

*Corresponding author: Tel.: +33-039-041-3553;
e-mail: srivastav@unistra.fr (Vinkle Srivastav)
[1]Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt and Tong Yu contributed equally and share co-first authorship.

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

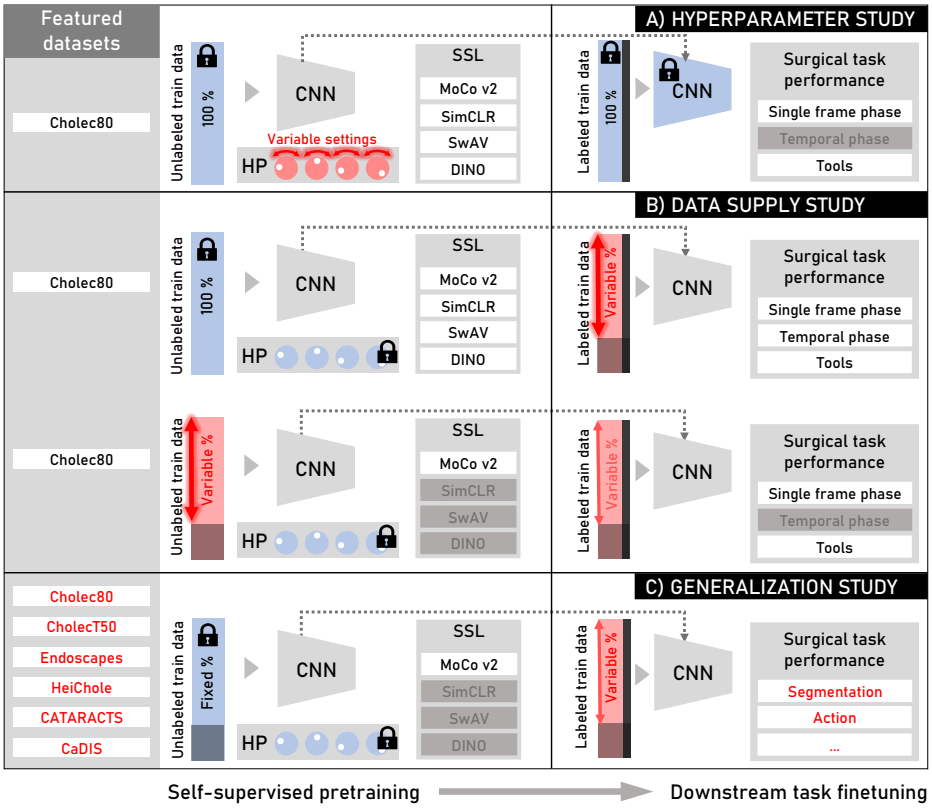

Fig. 1. Three stages of the study: (A) Hyperparameter study: Analyzing the influence of hyperparameters when adapting SSL methods to the surgical domain. (B) Data supply study: Evaluating the response of SSL methods to varying amounts of (1) labeled and (2) unlabeled data. (C) Generalization study: observing how well SSL generalizes to a much larger variety of surgical data and tasks.

(Yang and Kahrs, 2021) or robot kinematics (Sestini et al., 2021). State-of-the-art natural image SSL methods, with their advanced representational capabilities, have yet to be adequately demonstrated on surgical images.

However expanding SSL methods outside of natural images can be challenging, especially in a complex domain such as surgery. Most notably, heavy parameter tuning based on heuristics (Xiao et al., 2020) might be required. Robustness against large variations in domains and tasks also is not guaranteed; in-depth performance analysis has essentially been conducted on general computer vision datasets (Feichtenhofer et al., 2021a), most commonly Imagenet, which contains 14M images and over 1000 visually distinct classes. In contrast, Cholec80, one of the most prominent surgical computer vision datasets (Maier-Hein et al., 2017), contains 80 videos of procedures resulting in under 200k frames at 1fps. Only 7 classes of surgical phases and 7 classes of tools are featured; moreover, the visual evidence to distinguish them is highly sparse, especially for time-based tasks such as surgical phase recognition, a coarse-grained form of activity recognition. Further, since surgical videos can last up to several hours depicting a relatively stable scene, it is non-trivial to determine how existing SSL frameworks can best accommodate frames coming from the same procedure. Finally, these issues may be exacerbated by surgery-specific confounding factors such as smoke, bleeding, occlusions, or rapid tool movements. Such fundamental differences between natural and surgical image data motivate the need for a thorough study of SSL in the surgical domain.

The work presented here thoroughly addresses this need in three distinct steps (see Fig. 1). We select four SSL methods - MoCo v2 (Chen et al., 2020c), SimCLR (Chen et al., 2020b), SwAV (Caron et al., 2020), DINO (Caron et al., 2021) - suitably covering the state of the art in general computer vision, and extensively examine hyperparameter variations for each of them on Cholec80. We identify key differences with the natural image domain, highlighting hyperparameter tuning as a non-trivial and crucial element of SSL method transfer. In the second step, we set hyperparameters to their optimal values and test out the quality of the representations learned through each of these methods on two classic surgical downstream tasks: phase recognition and tool presence detection. Furthermore, we verify how these approaches respond to varying amounts of labeled and unlabeled data in a practical semi-supervised setting. Here, we show that these methods, while generic in design, achieve state-of-the-art performance for

both tasks and significantly mitigate the reliance on annotated data, adding up to 7.4% phase recognition F_1 score and 20.4% tool presence detection mAP. In the final step of the study, we extend our experiments to additional tasks and datasets: phase recognition & tool presence detection on HeiChole (Wagner et al., 2021), phase recognition & tool presence detection on CATARACTS (Al Hajj et al., 2019), action triplet recognition with CholecT50 (Nwoye et al., 2022b), semantic segmentation on Endoscapes (Alapatt et al., 2021), and 8 & 25 class semantic segmentation with CaDIS (Grammatikopoulou et al., 2021); thereby extensively covering the domain of surgical vision with SSL.

This paper's contributions are as follows:

1. Benchmarking of four state-of-the-art self-supervised learning methods (MoCo v2 (Chen et al., 2020c), SimCLR (Chen et al., 2020b), SwAV (Caron et al., 2020), and DINO (Caron et al., 2021)) in the surgical domain.
2. Thorough experimentation (~200 experiments, 7000 GPU hours) and analysis of different design settings - data augmentations, batch size, training duration, frame rate, and initialization - highlighting a need for and intuitions towards designing principled approaches for domain transfer of SSL methods.
3. In-depth analysis on the adaptation of these methods, originally developed using other datasets and tasks, to the surgical domain with a comprehensive set of evaluation protocols, spanning 10 surgical vision tasks in total performed on 6 datasets.
4. Extensive evaluation (~280 experiments, 2000 GPU hours) of the scalability of these methods to various amounts of labeled and unlabeled data through an exploration of both fully and semi-supervised settings.

2. Related Work

2.1. *Self-supervised representation learning in computer vision*

In the absence of external labels, SSL methods rely on the input image's intrinsic information to define a proxy loss to minimize. This artificial loss forces the model to learn rich vector representations of images, i.e. vectors in an embedding space with relative positions that meaningfully reflect the original visual content. The underlying expectation is that these representations are suitable for a wide range of useful downstream tasks.

The following paragraphs provide an overview of the various categories of SSL methods, tracing their evolution over the past few years. Here we focus on non-surgical visual tasks, considering mostly general computer vision works as well as a few others in medical image analysis.

Early heuristics-based methods. Early SSL approaches aimed to learn representations by training models to solve a simple handcrafted task with some degree of relevance to the target task (Kim et al., 2018). These included predicting spatial context (Doersch et al., 2015), image rotation (Gidaris et al., 2018), artificial classes based on geometric transformations (Dosovitskiy et al., 2014a), and image patch arrangement (Noroozi and Favaro, 2016). Similarly, other works proposed reconstructing image regions (Pathak et al., 2016) or colorization (Zhang et al., 2016, 2017). An exhaustive review of SSL methods based on pretext tasks is conducted in Jing and Tian (2020).

Contrastive methods. More recently, contrastive learning methods have emerged as an alternative to handcrafted heuristics. These methods place less emphasis on the nature of the pretext task, instead focusing on controlling the relative position of features in the embedding space. They rely on generating positive and negative pairs of samples, which are then passed to a discriminative loss function to generate a training signal.

Early works attempted to generate such samples from within a single image using image patches (Dosovitskiy et al., 2014b; Oord et al., 2018); however, these methods failed to take advantage of relationships between different images. Consequently, Wu et al. (2018) proposed the concept of a memory bank to store representations of many instances, which they leverage to impose an inter-instance discrimination objective. He et al. (2020) refined this idea with MoCo, using a momentum encoder rather than a memory bank to store representations, thereby enabling the sampling of many more instance pairs for the discrimination objective. An improved version with an additional projection head and more augmentations, MoCo v2, was later proposed by Chen et al. (2020c). Recently, Chen et al. (2020b) introduced SimCLR, a simpler framework outperforming many previous works (Oord et al., 2018; Bachman et al., 2019; Henaff, 2020; Tian et al., 2020; Misra and Maaten, 2020) by using aggressive data augmentations to generate 'positive pairs' for the discrimination objective.

Among SSL approaches, contrastive learning in particular has seen extensive use in research on medical image analysis in recent years. This form of pretraining has been employed to support many medical vision tasks: most commonly classification for diagnostic purposes (Chen et al., 2021; Ke et al., 2021; Yang et al., 2021; Xing et al., 2021; Dong and Voiculescu, 2021; Zhao and Yang, 2021; Huang et al., 2021; Dufumier et al., 2021), but also more complex tasks such as detection (Li et al., 2021; Tian et al., 2021; Lei et al., 2021), segmentation (Wu et al., 2021; Hu et al., 2021; Zeng et al., 2021; Boutillon et al., 2021; Zhou et al., 2021) and multimodal tasks combining text with vision (Liu et al., 2021; Jiao et al., 2020). Several imaging modalities are represented as well: MRI (Wu et al., 2021; Hu et al., 2021; Dufumier et al., 2021; Boutillon et al., 2021), CT (Yang et al., 2021; Lei et al., 2021; Zhou et al., 2021), X-Ray (Li et al., 2021; Liu et al., 2021) and ultrasound (Chen et al., 2021; Jiao et al., 2020).

Cluster-based and distillation-based methods. While contrastive methods have brought significant performance improvements, requiring positive and negative sampling during training can be impractical, and has pushed the community towards alternative approaches.

Self-supervised clustering methods (Caron et al., 2018; Asano et al., 2019; Caron et al., 2020; Grill et al., 2020a; Caron et al., 2021) provide another alternative to the pretext

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

task-based approach, focusing on clustering latent image representations in embedding space. Initially, Caron et. al. introduced DEEPCLUSTER (Caron et al., 2018), which adapted the k-means algorithm to assign clusters to images. Asano et al. (2019) showed reformulating cluster assignment as an optimal transport problem improves performance. SwAV (Caron et al., 2020) further improves on this by constraining augmented views of an image to have consistent cluster assignments.

Other works, based on distillation, bootstrap multiple neural networks in a teacher-student fashion to learn latent representations (Grill et al., 2020a). DINO (Caron et al., 2021) applies this bootstrapping approach with vision transformers, attaining state-of-the-art results.

Masked image modeling. Techniques based on concealing parts of images, as mentioned in our previous paragraph on heuristics-based methods, have existed in the computer vision community for several years: Pathak et al. (2016)'s image region reconstruction is one early example of masked image modeling (MIM). The emergence of Transformer models, however, led to a resurgence of MIM. Drawing inspiration from masked language modeling tasks for Transformers in natural language processing, recently published masked image modeling techniques view images as sequences of visual tokens, representing patches in a grid. A selection of tokens in the sequence is masked, then prompted for prediction by a Transformer employing attention on the sequence's tokens.

iGPT (Chen et al., 2020a) used a Transformer to predict individual pixels in images scaled down to low resolutions, while ViT (Dosovitskiy et al., 2021) predicted the mean colors of masked patches. BEiT (Bao et al., 2022), mc-BEiT (Li et al., 2022), and PeCo (Dong et al., 2021) learned to predict tokens produced by a VQ-VAE (Vector-Quantized Variational Auto-Encoder (van den Oord et al., 2017)) from masked patches. MaskFeat (Wei et al., 2022) studied a broad spectrum of feature types and proposed to regress Histograms of Oriented Gradients (HOG) for the masked content. MAE (He et al., 2022) and SimMIM (Xie et al., 2022) proceeded with direct regression on raw RGB pixel values.

Spatio-temporal methods. Parallel to static image methods presented in the previous paragraphs, research on SSL has explored video data through approaches tailored to spatio-temporal models. Most of them rely on spatio-temporal heuristics, with more emphasis on timing (Misra et al., 2016; Fernando et al., 2017; Lee et al., 2017; Xu et al., 2019; Wang et al., 2019; Jenni et al., 2020; Benaim et al., 2020) or appearance (Vondrick et al., 2018; Ahsan et al., 2019; Pathak et al., 2017; Kim et al., 2019; Diba et al., 2019). A few contrastive methods exist as well (Qian et al., 2021; Pan et al., 2021; Han et al., 2020). Recently, a large-scale study by Feichtenhofer et al. (2021b) adapted four single-frame SSL methods Chen et al. (2020b); He et al. (2020); Grill et al. (2020b); Caron et al. (2020) to video data and compared their performance.

Position of our work. Self-Supervised Learning is an intensely active research topic, with a large number of very distinct approaches proposed in recent years. For this reason, choosing an SSL method - especially for anything other than natural image data - is a complex problem: comparisons presented in SSL works can only cover a small selection of methods. More importantly, these comparisons are mainly conducted on natural image datasets such as the Imagenet dataset Deng et al. (2009); no reference point exists for surgical datasets, which are entirely different in terms of appearance. This is precisely the gap we fill with our work: we study how SSL adapts to surgical computer vision using a choice of methods that sufficiently span the state-of-the-art for static images with methods based on contrastive learning, clustering, and distillation. Masked Image Modeling methods have not been selected since the patch division process that makes those suitable for Transformers would first need to be ported to the more classical architecture of ResNet50 (retained due to its status as the standard for SSL). This port alone would require extensive and dedicated experimentation. Spatio-temporal models, while potentially relevant for future studies, are also omitted here due to challenging and radically different temporal modeling requirements in the surgical domain: commonly used natural video datasets in SSL (Carreira and Zisserman, 2017; Soomro et al., 2012; Kuehne et al., 2011) contain short clips of a single action, contrasting heavily with full recordings of surgical interventions.

2.2. *Surgical computer vision.*

General computer vision focuses on natural images with scenes and items from everyday life. In contrast, surgical computer vision aims at identifying surgical activities and objects with varying degrees of detail. Early work in the field focused on automatically recognizing surgical workflow at the coarsest level through two fundamental tasks: phase recognition and tool presence detection. These highly specialized visual tasks prompted developments in terms of methodology separately from the rest of computer vision, which we cover in the next paragraphs.

Full supervision. Initial efforts in surgical computer vision involved phase recognition based on handcrafted features (Padoy et al., 2012; Blum et al., 2010). Deep learning was first introduced to the field by Twinanda et al. (2016b) and Dergachyova et al. (2016), replacing handcrafted features with embeddings extracted by convolutional neural networks; Twinanda et al. (2016b) in particular introduced the *Cholec80* dataset, containing 80 videos of cholecystectomy annotated with surgical phases and tool presence labels. This dataset has since remained as one of the surgical computer vision community's main datasets (Maier-Hein et al., 2017), appearing in most works mentioned in this paragraph. With surgical workflow and continuity of surgical actions playing a major role in these tasks, spatio-temporal models quickly emerged, outperforming single-frame models by a wide margin. Twinanda et al. (2016a) employed combinations of CNNs and LSTMs for surgical phase recognition and tool presence detection. Since then, increasingly refined spatio-temporal architectures have been proposed to better model the tasks (Jin et al., 2018, 2020; Czempiel et al., 2020; Jin et al., 2021;

Czempiel et al., 2021). Recently, Rivoir et al. (2022) studied end-to-end spatio-temporal models and the effect of Batch Normalization on the success of these models. Outside of these examples, a more comprehensive overview of surgical phase recognition approaches is provided in a survey by Garrow et al. (2021). For recognizing tools in cataract surgery, Al Hajj et al. (2018) proposed combinations of CNNs and RNNs with boosting.

Self-supervision in surgery. Self-supervision is still in the very early stages of research within surgical computer vision. While SSL methods in general computer vision have evolved towards methods such as contrastive learning, clustering or distillation (Section 2), self-supervision on surgical data is still mostly limited to heuristics; for instance, Ross et al. (2018) uses a colorization pretext task. Furthermore, the self-supervised tasks seen in surgery generally involve external information: da Costa Rocha et al. (2019); Sestini et al. (2021) incorporate robot kinematics. Yengera et al. (2018) rely on remaining surgery duration estimation as the pretext task to improve surgical phase recognition on Cholec80. The only existing examples of contrastive learning add external information as well: Bodenstedt et al. (2017) used a frame sorting task; later, Funke et al. (2018) introduced a method named second-order temporal coherence. In both cases, comparisons between frames are driven by time (i.e. relative positions of frames inside of a video) instead of their actual content.

Position of our work. Current research on surgical computer vision heavily leans towards fully supervised methods, which require large amounts of data to be annotated with clinical expertise. For improved scalability, a few approaches involving self-supervision have been developed. These approaches, however, heavily rely on heuristics and external information; as such, they lag behind general SSL, which has expanded to a larger spectrum of methods in recent years, all purely based on pixel data. Our work targets this deficit by bringing recently proposed SSL methods to surgery and adapting them to this particular domain. Since single-frame feature extractors play a fundamental role in state-of-the-art spatio-temporal models in surgical computer vision, examining SSL methods designed for static images is an obligatory first step, which is the focus of this study.

3. Methodology

We first establish the setting of this study by introducing the relevant surgical data and tasks, followed by our selection of SSL methods. We then outline our experiments; three main stages are defined as shown in Fig. 1, the *hyperparameter study* (A), the *data supply study* (B) and the *generalization experiments* (C). Stages A and B each examine in detail the reaction of SSL in the surgical domain to a different factor, respectively parameterization and available data quantities. Stage C is an extension of our experiments to a much larger variety of datasets and tasks. Implementation details for each stage of this study are available in the supplementary material.

3.1. Surgical data & surgical tasks

Cholec80. Since its introduction by Twinanda et al. (2016b), the Cholec80 dataset has been the foundation for many studies in surgical computer vision; we, therefore, use it here for our SSL benchmark. This dataset contains 80 videos of complete laparoscopic cholecystectomy procedures, recorded at 25 frames per second with a resolution of 854 × 480 or 1920 × 1080. The average video duration is 38 minutes with 16 minutes of standard deviation, indicating a high degree of heterogeneity.

The two tasks used as downstream tasks are *tool presence detection* and *surgical phase recognition*, mirroring the *object detection* and *action recognition* tasks of general computer vision, respectively.

Tool presence detection is a multi-class, multi-label classification problem aimed at identifying all the surgical tools appearing in a given frame (Twinanda et al., 2016b; Nwoye et al., 2019; Al Hajj et al., 2018). It goes beyond image-level classification as zero, one, or several types of tools can be detected in one surgical image frame at the same time. 7 tools are featured, as described in Fig. 2.

Surgical phase recognition entails classifying every frame of a recorded surgical procedure based on the activity being performed. This is a challenging task since important tools or anatomical parts often exit the field of view; as a result, useful visual indicators for making predictions tend to be quite sparse. Each procedure is decomposed into up to 7 phases described in Fig. 3.

CHOLEC80 TOOLS			
Name	Function		Occurrences per video
Grasper	Hold or move anatomy		1282±1669
Bipolar	Coagulate, hold or move anatomy with a pair of electrodes		111±106
Hook	Dissect tissue or coagulate with an electrode		1289±672
Clipper	Ligate using clips		41±31
Scissors	Perform cuts		75±48
Irrigator	Project water, aspirate fluids		123±147
Specimen bag	Carry gallbladder		143±84

Fig. 2. Tools featured in the Cholec80 dataset.

Additional data & tasks. While experiments featured in this work mostly focus on Cholec80 due to its prevalence in the community, a later stage of our study looks at other interesting datasets and surgical tasks. The digest of all datasets and tasks are presented in Fig. 6.

HeiChole. The HeiChole[2] (Wagner et al., 2021) dataset,

[2] https://www.synapse.org/#!Synapse:syn18824884/wiki/

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

CHOLEC80 PHASES			
Name	Description		Duration
Preparation	Exposure of gallbladder by removal of surrounding tissue		1.8±1.7 min
Calot triangle dissection	Exposure of the base of the liver bed by dissecting the gallbladder neck		15.6±11.1 min
Clipping & cutting	Application of clips to the cystic duct, cutting of cystic duct		2.9±2.1 min
Gallbladder dissection	Dissection of gallbladder from the liver bed		12.2±8.9 min
Gallbladder packaging	Insertion of dissected gallbladder into specimen bag		1.6±0.8 min
Cleaning & coagulation	Coagulation of the liver bed and cleanup using the irrigator		3.0±2.6 min
Gallbladder extraction	Extraction of the gallbladder through the umbilical trocar		1.4±1.2 min

Fig. 3. Phases featured in the Cholec80 dataset.

introduced as part of the EndoVis 2019 challenge, consists of 33 video recordings of cholecystectomy surgeries from three different hospitals. The training set, consisting of 24 videos, is publicly available while a test set of 9 videos is privately held for evaluation. The complete dataset contains frame-wise annotations of surgical phase and tool presence. Each procedure is segmented into 7 phases and could feature up to 7 tools. The description of all the phases and tools is presented in Wagner et al. (2021).

CATARACTS. The CATARACTS dataset, introduced as part of the Challenge on Automatic Tool Annotation for cataRACT Surgery (CATARACTS)[3] in 2017, is another popular dataset in the surgical vision community. The dataset consists of 50 recordings of cataract surgical procedures. In a recent edition of the challenge[4] (Al Hajj et al., 2019), the dataset was fully annotated for both tool presence detection and surgical activity recognition (step) tasks. In total, there are 19 steps and 21 different tool classes. We use the same splits as the CATARACTS 2020 challenge where the dataset was separated into 25, 5, and 20 videos corresponding to a train, validation, and test set, respectively.

CholecT50. CholecT50 is a video dataset of laparoscopic cholecystectomy surgery introduced by Nwoye et al. (2022b) to enable research on fine-grained action recognition. A collection of 50 videos, of which 45 videos are from the Cholec80 dataset and an additional 5 videos from an in-house dataset for cholecystectomy surgery, are fully annotated with action triplet information in the form of ⟨*instrument, verb, target*⟩. A total of 100 actions triplet classes are defined by Nwoye et al. (2022b) as various combinations of 6 instruments, 10 verbs, and 15 targets. The dataset is split into 45 videos for training and 5 videos for testing, following the split used in the CholecTriplet2021 Challenge [5].

Endoscapes. Introduced by Alapatt et al. (2021), Endoscapes is a dataset comprised of 2208 frames selected at regular intervals (every 30 seconds) from 201 laparoscopic cholecystectomy videos with pixel-wise annotations for the task of semantic segmentation. A total of 29 semantic classes are defined in Alapatt et al. (2021) with 6 anatomy classes, 19 instrument classes, and 4 other miscellaneous classes. We follow the same data splits of Alapatt et al. (2021) in all our experiments. **CaDIS.** CaDIS (Grammatikopoulou et al., 2021) is a semantic segmentation dataset for cataract surgery. The dataset consists of 4670 images extracted extending part of the CATARACTS dataset with pixel-level annotations for 36 classes (29 surgical instrument classes, 4 anatomy classes, and 3 miscellaneous classes). The 4670 images are split into train, validation, and test sets comprising 3550, 534, and 586 images, respectively. Out of the three different evaluation tasks, representing increasing degrees of granularity, we consider the two extremes for evaluation in this study. Task I aims at differentiating anatomy and instruments in each frame and hence consists of 8 semantic classes: 4 classes for anatomical structures, 1 class for all instruments, and 3 classes for all other objects appearing in the images. Task III, on the other hand, focuses on more detailed instrument classification by representing each instrument type and instrument tips as separate classes totaling 25 classes.

3.2. Selected SSL methods

As shown in Section 2, general computer vision offers a wide range of SSL methods. In order to adequately represent the current state of the art, we select a total of four SSL methods: two contrastive (SimCLR (Chen et al., 2020b), MoCo v2 (He et al., 2020; Chen et al., 2020c)), one distillation-based (DINO (Caron et al., 2020)), and one clustering-based (SwAV (Caron et al., 2020)), see Fig. 4.

Several studies on unsupervised visual representation have proposed approaches based on contrastive learning (Hadsell et al., 2006; Wu et al., 2018; Van den Oord et al., 2018; Hjelm et al., 2018; Zhuang et al., 2019; Henaff, 2020; Tian et al., 2020; Bachman et al., 2019), with the core idea being to maximize the representational similarity for pairs of positive samples and dissimilarity for pairs of negative samples. A key component of these methods is mining positive and negative samples in a batch without explicit labels. A common approach in these methods is, for each image, to consider its augmentations as a corresponding positive sample, and other images as corresponding negative samples. The positive and the negative samples are passed through a base encoder to obtain the corresponding positive (x, x^+) and negative (x^-) embeddings. The InfoNCE loss (Oord et al., 2018) commonly used in contrastive methods is defined as follows:

$$L_{contrastive} = \mathbb{E}_{x,x^+,x^-}\left[-\log \frac{e^{x\cdot x^+/\tau}}{e^{x\cdot x^+/\tau} + (\sum_{k=1}^{K} e^{x\cdot x^-/\tau})}\right], \quad (1)$$

591922

[3] https://cataracts.grand-challenge.org/

[4] https://www.synapse.org/#!Synapse:syn21680292/wiki/601561

[5] https://cholectriplet2021.grand-challenge.org/

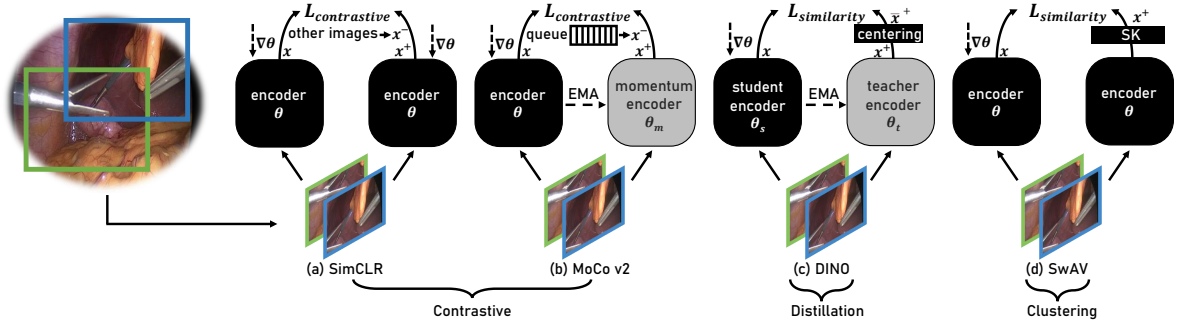

Fig. 4. We study four SSL methods from three categories: contrastive (SimCLR (Chen et al., 2020b) and MoCo v2 (He et al., 2020; Chen et al., 2020c)), distillation-based (DINO (Caron et al., 2021)), and clustering-based (SwAV (Caron et al., 2020)). SimCLR and MoCo v2, as contrastive methods, use embeddings from other images or a queue to generate negative embeddings (x^-), respectively. MoCo v2 and DINO use an explicit momentum encoder whose weights are updated using an exponential moving average (EMA). $\nabla\theta$ are the gradients of the encoder's weights θ, computed using a contrastive loss ($L_{contrastive}$) for SimCLR and MoCo v2 and a similarity loss ($L_{similarity}$) for DINO and SwAV. DINO uses a centering operation, and SwAV uses a non-differentiable Sinkhorn-Knop (SK) transform (Cuturi, 2013) to avoid mode collapse in the absence of negative embeddings.

where τ is a temperature hyperparameter for scaling the embeddings. The negative samples are required in contrastive methods to avoid model collapse to an identity solution. Each of the following four selected SSL methods works on similar principles with a few modifications.

SimCLR (Chen et al., 2020b) considers the other images from a batch as negative samples and passes them through the encoder to obtain the negative embeddings (x^-) to compute the contrastive loss, $L_{contrastive}$, using equation (1).

MoCo v2 He et al. (2020) introduced MoCo, employing a large memory queue to store negative embeddings x^-. This queue allows decoupling the dictionary size from the mini-batch size, in order to perform well even with smaller batch sizes. Furthermore, since the queue contains embeddings from different mini-batches, a momentum encoder is used to enforce consistency across different mini-batches. The weights of the momentum encoder (θ_m) are updated using an exponential moving average (EMA) of the weights of the encoder (θ): $\theta_m = \lambda\theta_m + (1-\lambda)\theta$, where λ is a decay parameter. MoCo v2 (Chen et al., 2020c) refines this design using an additional projection head and more augmentations.

DINO (Caron et al., 2021), inspired by BYOL (Grill et al., 2020b), uses a teacher-student approach in a knowledge-distillation framework (Hinton et al., 2015). The student encoder, parameterized by θ_s, and the teacher encoder, parameterized by θ_t, are used to generate two positive embeddings, x and x^+, respectively. Similar to MoCo v2, the weights of the teacher encoder are updated using EMA. However, DINO also removes the dependency on negative samples; in the absence of negative embeddings, this method avoids *model collapse* using a *centering* operation. This operation first computes the centers of the positive embeddings using EMA, $c = \lambda_c c + (1-\lambda_c)\frac{1}{B}\sum_{i=1}^{B} x_i^+$, then subtracts the centers c from the positive embeddings to compute the mean-centered positive embeddings, $\bar{x}^+ = x^+ - c$. Here, B is a batch dimension and λ_c is a centering decay parameter. The similarity loss

$$L_{similarity} = -\sum \text{softmax}(x/\tau_s)\log(\text{softmax}(\bar{x}^+/\tau_t)) \quad (2)$$

is computed as a cross-entropy loss between the reference positive embedding, x, and mean-centered positive embeddings, $\bar{x}^+$. The softmax() function normalizes embeddings that are scaled differently using temperature parameters τ_s and τ_t for the student and teacher encoders, respectively.

SwAV (Caron et al., 2020) circumvents the need for negative embeddings by first transforming the positive embedding pair, x and x^+, to learned prototype embeddings, $\bar{x}$ and $\bar{x}^+$ and then performing online clustering of the learned prototype embeddings using the Sinkhorn-Knopp (SK) algorithm (Cuturi, 2013). The SwAV similarity loss is

$$L_{similarity} = \mathcal{D}_{KL}(\bar{x} \parallel \text{SK}(\bar{x}^+)), \quad (3)$$

where $\mathcal{D}_{KL}$ is the Kullback-Leibler divergence.

3.3. Hyperparameter study design

In the hyperparameter study (Fig. 1, A), we aim to better understand the sensitivity of each SSL method to hyperparameter variations and establish a set of **recommended** values that will later serve in practical use cases of semi-supervised learning, as part of the data supply study (Fig. 1, B). To this end, we select a subset of 5 critical hyperparameters:

- Type of augmentation
- Batch size
- Epochs
- Sampling rate
- Type of initialization

We then carefully analyze the influence of all 5 on the model performance, for the tasks of phase recognition and tool presence detection on the Cholec80 dataset. Each of those 5 hyperparameters defines a group of experiments, where the relevant hyperparameter varies while others are set to the default values shown in Table 1. For each value of that hyperparameter, 4 models are trained - one for each selected

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

Table 1. Observed SSL hyperparameters. Defaults are used in the hyperparameter study. Recommended values (best overall performance in the hyperparameter study) are used in the data supply study.

		Defaults	Recommended
Augmentations	Multi-Crop	8	2
	Color	On	On
	Geometric	On	On
	Strong-color	Off	Off
Batch size		512	256
Epochs		300	300
Sampling rate		1	5
Initialization		Scratch	Imagenet fully supervised

SSL method. Linear evaluation is then performed on the validation set, i.e. by training a linear classifier added on top of the frozen backbone layers, for tool and phase tasks separately. This validation protocol, commonly used in SSL (Feichtenhofer et al., 2021a), verifies here how well each method, for that particular hyperparameter value, maps frames to linearly separable vector representations that are consistent in terms of phase and tool content. Details for each experiment group are provided in the following paragraphs.

Augmentations. Data augmentation is a crucial aspect of SSL methods (Chen et al., 2020b): learning persistent feature representations between different *views* of the same image (i.e. between different augmented versions of the original image), is the implicit task that SSL methods leverage in order to produce powerful representations of unlabeled data. Hence, it is imperative to understand the impact of this parameter when shifting to different domains and tasks. While an exhaustive search of augmentations is beyond the scope of this study [6], we decided to focus on broad categories of commonly used augmentation techniques to train SSL methods (Caron et al., 2021; Chen et al., 2020b; He et al., 2020), defined here as *Color*, *Geometric*, *Strong-color* and *Multi-Crop*. Fig. 5 provides a description for each category.

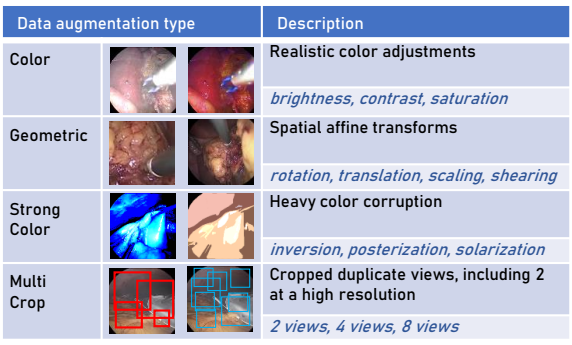

Fig. 5. Data augmentation types involved in the hyperparameter study

[6] Pretraining a ResNet-50 using SSL with a single hyperparameter setting given our experimental design demands approximately 40 GPU hours using 4 NVIDIA V100s on average across considered methods.

All the mentioned augmentations are randomized during training (Cubuk et al., 2020a); the randomization process follows the implementation of Goyal et al. (2021).

Multi-Crop is set to 2, 4, or 8 crops with 2 crops always sampled at a high resolution following Caron et al. (2020). Each of the other 3 augmentation types is either *on* or *off*. Considering all the possible combinations, we examine a total of $3 * 2^3 = 24$ configurations for augmentations.

Batch size. Batch size is a crucial hyperparameter in SSL methods: SimCLR (Chen et al., 2020b) established a positive correlation between performance and batch size attributed to the size of the pool of negative samples to draw from during training. The other 3 approaches have presented the ability to better function with smaller batches as an advantage, cutting down memory requirements.

To examine these claims, we use batches of sizes 128, 256, 512, and 1024.

Epochs. Previous studies have shown that training time could largely impact SSL performance. Given this, we investigate the impact of training time by training each SSL method for 50, 100, 200, and 300 epochs.

Sampling rate. While the SSL methods we test are designed for still images, we can apply them to video inputs by simply extracting individual frames from each video. A key consideration when doing so is the frame sampling rate, as this can affect the relative homogeneity among various input images. In this aspect, surgical videos pose a particularly interesting technical setting, as they tend to provide a stable context, and the only changes across frames, even for several minutes of video, are manipulations of organs and medical tools in the field of view. Consequently, while increasing the number of frames sampled per second dramatically increases the available training data, it is unclear whether this additional data would be beneficial for SSL methods.

We experiment with sampling videos at 0.1, 0.33, 0.5, 1, 3, and 5 frames per second (fps).

3.4. Data supply study design

In contrast with the previous section, the data supply study (Fig. 1, B) operates with a completely fixed set of recommended hyperparameters (Table 1), suitable for examining our chosen SSL methods in practical semi-supervision use cases: instead of freezing the backbone after self-supervised training, here we finetune it with phase or tool annotations in conjunction with a linear classifier. For phase recognition, we also observe the performance obtained by adding a temporal model (TCN, Czempiel et al. (2020)) after this step and finetuning it separately as well: this provides a strong point of comparison against the state of the art, while also gauging the representations learned through SSL when used in a temporal context.

Labeled data supply. We first focus on labeled data only. Performance with respect to annotated data availability (Fig. 1, B1) is examined in three settings, with supervised finetuning performed after SSL on 40 videos (100% of the entire Cholec80 training set), 10 videos (25%), or 5 videos (12.5%) of the full data. To mitigate the effect of outliers,

experiments for the last two settings are replicated on 3 randomly selected sets of videos. In all these configurations, the same 40 unlabeled videos are used for self-supervised pretraining.

Unlabeled data supply. In addition to this core set of experiments focusing exclusively on varying labeled data, we select one SSL method - MoCo v2 - and examine how it reacts to changes in the amount of unlabeled data (Fig. 1, B2) used for self-supervised training: from 1 to 10, 20, 40 and finally 80 unlabeled videos. Results are reported for varying numbers of labeled videos used for finetuning.

3.5. Generalization study

Experiments conducted up to this point feature the Cholec80 dataset with two tasks - phase recognition and tool detection - representing only a small portion of the variability of datasets used in surgical data science literature (Maier-Hein et al., 2022). In order to determine how well SSL generalizes to entirely different situations within surgery, we provide in this final stage a set of complementary experiments of a previously selected SSL method - MoCo v2 - inspecting its behavior across a total of 8 tasks across 5 different surgical datasets: HeiChole (Wagner et al., 2021), CATARACTS (Zisimopoulos et al., 2018), CholecT50 (Nwoye et al., 2019), Endoscapes (Alapatt et al., 2021), and CaDIS (Grammatikopoulou et al., 2021). Here the scope of the study is expanded by a considerable amount in several aspects. First, we study the effect of the SSL methods on the same surgical procedure and tasks but on diverse clinical centers, with surgical data sourced from 3 German hospitals (HeiChole). Next, we investigate another type of minimally invasive surgery, i.e., cataract, through the CATARACTS dataset, offering a radically different visual appearance from cholecystectomy. Here again, we consider similar downstream tasks of surgical activity (step) recognition and tool presence detection. We further extend our analysis of SSL methods on yet another task, surgical action triplet recognition, on the recently released CholecT50 dataset. We add surgical scene segmentation as well with the Endoscapes dataset. Finally, we conclude the generalization study by analyzing the SSL methods on another surgical procedure and task with the CaDIS dataset for scene segmentation in cataract surgery. A visual summary of the different dataset characteristics is shown in Fig. 6.

Additional data & tasks

Dataset		Surgery	Video source	Tasks	# of classes
HeiChole		Cholecystectomy	Heidelberg	Phase	7
				Tool	7
CATARACTS		Cataract	Brest	Step	19
				Tool	21
CholecT50		Cholecystectomy	Strasbourg	Action	100
Endoscapes		Cholecystectomy	Strasbourg	Segmentation	29
CaDIS		Cataract	Brest	Segmentation	8
				Segmentation	25

Fig. 6. Data featured in the generalization experiments.

4. Results

4.1. Dataset Splits and Evaluation Metrics

In all our experiments, following previous literature (Czempiel et al., 2020; Jin et al., 2018; Twinanda et al., 2016b; Czempiel et al., 2021), we use 40, 8, and 32 videos from Cholec80 as our total available pool of training videos, our validation set, and our test set, respectively.

In the hyperparameter study, we perform SSL pretraining on the entire pool of 40 training videos and report the results on the validation set.

In the data supply study, we further conduct semi-supervised experiments with 5 videos (12.5% of Cholec80 training set) and 10 videos (25% of Cholec80 training set) of annotations, for which we employ two different sampling strategies. For the comparison with external methods (Table 6), we use the predefined dataset split introduced in Shi et al. (2021) as a sampling strategy to enable fair comparisons. However, for the remainder of our experiments (see Tables 3, 4, 5, and Figures 13, 14), we either make use of established training splits (Twinanda et al., 2016b) for larger data settings (40, 80 training videos), employ a stratified random sampling approach or random uniform sampling when stratifying is infeasible (1 training video). In each case when randomly sampling, we sample three separate subset splits of the training videos, evaluate model performance on each split, and report the mean and standard deviation across splits. Doing so alleviates selection bias and allows for sound comparisons across methods and experimental settings. Indeed, we find that the variance in performance across dataset splits, particularly in the low-data settings, can surpass performance differences across methods, highlighting the need to sample multiple splits.

For all phase and step recognition experiments, with the exception of the external comparison (Table 6), we report per-video F1 Score, computed by averaging across each video's F1 score. In these tables, the standard deviation is presented across the sampled splits. Meanwhile, for the external comparison, we report a *relaxed boundary* per-video F_1 Score, originally introduced in the m2cai16-workflow challenge [7] and used by Shi et al. (2021), to enable a fair comparison. The relaxed boundary metric introduces a 10 second 'relaxed' period centered around each ground truth phase transition; during these periods, the two consecutive phases are considered to be correct classifications (e.g. phase 4 and phase 5 are both accurate classifications in the 10 seconds before and after the transition from phase 4 to 5). Consequently, the relaxed boundary metric results in higher scores across methods.

For all tool presence detection experiments, we compute mAP across all considered frames and in all the presented tables the standard deviation is calculated across splits. Action

[7] http://camma.u-strasbg.fr/m2cai2016/index.php/program-challenge/

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

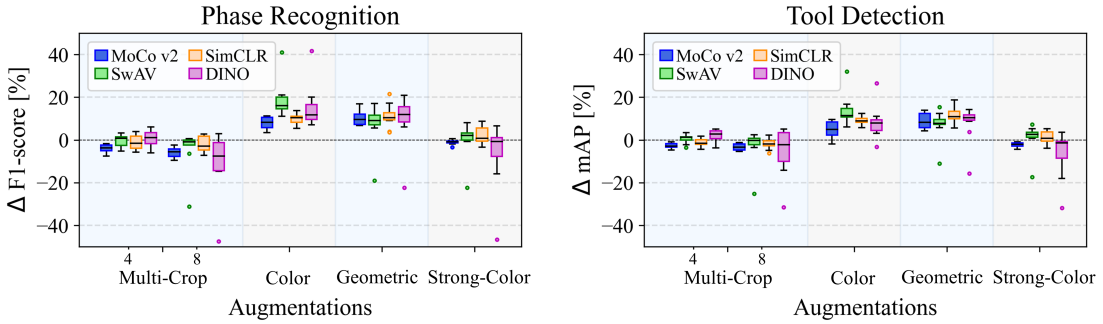

Fig. 7. Performance of each method on Cholec80 varying the augmentation strategy for self-supervised pretraining. For each method and category of augmentations, we show a boxplot with the change in performance from the default no-augmentation setting (using 2 crops for *Multi-Crop*), by enabling that category of augmentation (using 4 or 8 crops for *Multi-Crop*). The boxplot whiskers were set to 1.5 times the interquartile range beyond the first and third quartile; settings outside of this margin were defined as outliers and plotted as dots. Results were obtained using linear evaluation on the validation set. Left: F_1-score for phase recognition. Right: mAP for tool presence detection.

triplet recognition performance on the CholecT50 dataset is measured using mAP over the 100 valid triplet classes. Segmentation tasks featured in the generalization experiments are all evaluated using F_1 score.

4.2. *Hyperparameter study*

We present here the impact of hyperparameters variations on the quality of the representations learned by the SSL methods we selected, following the setup described in Section 3.3 [8].

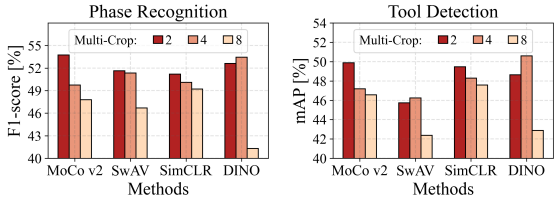

Fig. 8. Performance of each method on Cholec80 varying the Multi-Crop augmentation strategy for self-supervised pretraining: 2,4 or 8 crops (2 high-resolution crops, remaining low resolution). Results were obtained using linear evaluation on the validation set. Left: F_1-score for phase recognition. Right: mAP for tool presence detection.

t

Augmentations. In order to evaluate the impact of each of the four augmentation categories, we show the improvement introduced by the presence of each category across all the considered experiments for each SSL method. For every augmentation category, we examine the change in performance - ΔF_1 and ΔmAP - caused by toggling it *on* (for *Multi-Crop*, by switching it from 2 to either 4 or 8). To this end, in Fig. 7, we plot the following set of samples for the *Multi-Crop* (4 and 8 crops - **MC4** and **MC8**), *Color* (**C**), *Geometric* (**G**) and *Strong-Color* (**S**) augmentation experiments, respectively:

$$\begin{aligned}
\textbf{MC8} &= \{(mc_{\mathbf{8}}\ c_i g_j s_k - mc_{\mathbf{2}} c_i g_j s_k)_{i=\{1,0\},j=\{1,0\},k=\{1,0\}}\},\\
\textbf{MC4} &= \{(mc_{\mathbf{4}}\ c_i g_j s_k - mc_{\mathbf{2}} c_i g_j s_k)_{i=\{1,0\},j=\{1,0\},k=\{1,0\}}\},\\
\textbf{C} &= \{(mc_i c_{\mathbf{1}} g_j s_k - mc_i c_{\mathbf{0}} g_j s_k)_{i=\{2,4,8\},j=\{1,0\},k=\{1,0\}}\},\\
\textbf{G} &= \{(mc_i c_j g_{\mathbf{1}} s_k - mc_i c_j g_{\mathbf{0}} s_k)_{i=\{2,4,8\},j=\{1,0\},k=\{1,0\}}\},\\
\textbf{S} &= \{(mc_i c_j g_k s_{\mathbf{1}} - mc_i c_j g_k s_{\mathbf{0}})_{i=\{2,4,8\},j=\{1,0\},k=\{1,0\}}\},
\end{aligned} \tag{4}$$

where *mc* is *Multi-Crop* augmentation and can take the values 2,4,8; *c*, *g*, *s* are, respectively, *Color*, *Geometric* and *Strong-Color* augmentations, which can either be toggled *on* (1) or *off* (0). For each augmentation setting, statistics for ΔF_1 and ΔmAP are collected and represented as boxplots. The average performance for each *Multi-Crop* setting is also shown separately in Fig. 8.

Experimental results for phase recognition and tool presence detection, shown in Fig. 7, demonstrate the clear impact that augmentation strategies have on the quality of the learned representations, consistent across methods and tasks. We make three main observations:

(1) In general, increasing the number of low-resolution views on *Multi-Crop* negatively impacts performance. From 2 crops for MoCo v2, switching to 4 crops cuts down phase recognition F_1 by 3.5%; switching to 8 cuts it down by 4.5%. This represents an important deviation from typical results in the natural image domain, where additional low-resolution views in *Multi-Crop* generally positively correlated with improved performance (Caron et al., 2020, 2021). A possible explanation may be the weaker value of ensuring 'local-to-global' feature invariance in the surgical domain; in surgical phase recognition, for example, discriminative cues may be scattered in the entire image, and be significant only if considered as a whole: in light of this, forcing 'local-to-global' invariant features may be challenging, or even undesirable in this domain.

[8] GPU training presents some non-determinism that is not trivial to avoid. Because performing several reruns of every experiment in the hyperparameter study would be computationally impractical, we do so for one method selected at random and present the standard deviation when performing linear evaluation for both downstream tasks in order to contextualize our results. The standard deviation across 5 reruns for this selection for phase recognition and tool presence detection is 0.7 % F1 and 0.7 % mAP, respectively.

(2) The *Color* augmentation consistently and significantly improves performance. This is generally analogous to results on the natural image domain (Feichtenhofer et al., 2021a): as pointed out in (Chen et al., 2020b), augmentations like *Multi-Crop* and *Geometric* mostly preserve the original color distribution, leaving this as an easy shortcut for the network to solve the predictive task; the *Color* augmentation is, therefore, an important factor in learning meaningful representations.

(3) DINO is the method most affected by the specific choice of augmentation; in particular, representation quality dramatically drops when both *Multi-Crop* and *Strong-Color* augmentations are used; a possible explanation may derive from the general observation on *Multi-Crop* made previously: compared to the other methods, DINO explicitly enforces the 'local-to-global' feature invariance by passing all views to the student, but only global *views* to the teacher. While this task is intrinsically difficult in the surgical domain, for the previously discussed reasons, it may be made even more challenging by the presence of the *Strong-Color* augmentation, leading to unreliable feature representations.

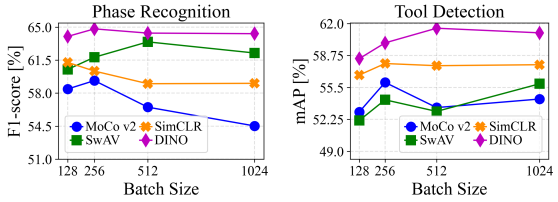

Fig. 9. Performance of each method on Cholec80 varying the batch size used for self-supervised pretraining. Results were obtained using linear evaluation on the validation set. Left: F_1-score for phase recognition. Right: mAP for tool presence detection.

Batch size. Overall, larger batch sizes do not improve feature quality. Clear improvements are only perceivable between 128 and 256 (up to 4.8% F_1 for phase recognition, 5.6% mAP for tool detection) across all tasks and methods - except for phase recognition with SimCLR. Results for 256 and above, however, generally contradict claims from other SSL works (Chen et al., 2020b; Caron et al., 2020, 2021), especially on the phase recognition task (Fig. 9): from 256 to 1024, MoCo v2's F_1 score drops by 5.5%. No clear positive impact of increasing batch size past 256 can be seen on tool presence detection either (Fig. 9).

This inconsistency with results obtained on natural images is possibly due to differences in data scale since Cholec80 (at 1 fps: $\sim 10^5$ samples, 7 classes) is far smaller than ImageNet ($> 10^6$ samples, 10^3 classes). During training, batches are therefore sampled under completely different conditions; since SSL methods, in the absence of labels, rely heavily on negative and positive samples to separate classes, this can affect the final performance.

In the literature, one documented adverse effect of larger batches in SSL is shown by Chen et al. (2020b) on SimCLR, when the batch size is pushed up to high values (>2048). A scaled-back version of this phenomenon might be at play here.

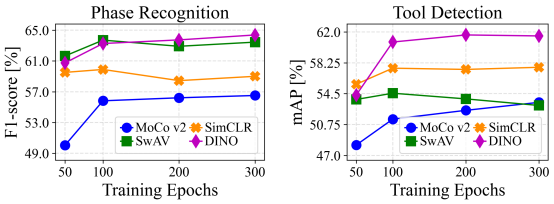

Fig. 10. Performance of each method on Cholec80 varying the number of epochs used for self-supervised pretraining. Results were obtained using linear evaluation on the validation set. Left: F_1-score for phase recognition. Right: mAP for tool presence detection.

Epochs. Overall, phase recognition and tool presence detection performance (Fig. 10) tends to saturate as epochs increase, with nuances from one SSL method to another. SwAV and SimCLR in particular clearly peak earlier than the other two methods at 100 epochs, losing up to 2% phase recognition F_1 and 2% tool presence detection mAP afterward. In contrast, MoCo v2 and DINO improve over the entire 300-epoch training period, with, nonetheless, a noticeable slowdown after 100 epochs.

This disparity could be a result of including a momentum encoder (used by both MoCo v2 and DINO). The momentum encoder enables a greater diversity in pairs of latent vectors generated by the network backbone during training: in MoCo v2, via a greater set of negative samples to choose from, and in DINO, via the teacher network incorporating context from a wider variety of samples. Consequently, longer training may allow models to learn more robust representations.

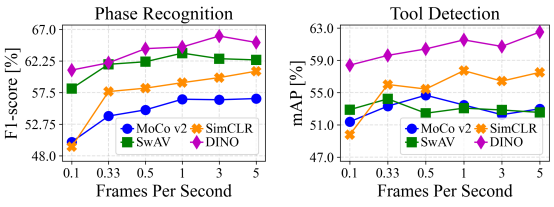

Fig. 11. Performance of each method on Cholec80 varying the Frames Per Second for self-supervised pretraining. Results were obtained using linear evaluation on the validation set. Left: F_1-score for phase recognition. Right: mAP for tool presence detection.

Sampling rate. As previously stated, surgical videos pose a particularly interesting technical setting for SSL research in general because surgical videos often provide a very stable context while the anatomy in the scene is manipulated. While increasing the number of frames sampled per second could dramatically expand the available training data, performance might not increase due to redundancy. Indeed, with the 5 sampling rates examined here, we observe marginal utility in sampling frames beyond a certain frequency. For both tasks, when sampling frames at over 1 fps, we observe no consistent improvement across methods or tasks when training

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

Table 2. The average results across methods are presented for phase recognition, tool presence detection and the average across both tasks (Selection metric). For each individual ablation, results are presented in descending order of performance according to the Selection metric. The Setting column refers to the value of the parameter being ablated, while all other settings are kept to the default values specified in Table 1. For the augmentation ablation, we use the following notations: MC - Multi-Crop, C - Color, G - Geometric, S - Strong-color; for the MC setting columns, we specify the total number of crops used (including 2 high-resolution crops) and for the S, G, and C setting columns, we specify whether those augmentation categories were included or "on".

Ablation	Setting	Selection metric	Phase (F1)	Tool (mAP)
Sampling rate	5.0	58.8	61.2	56.4
	1.0	58.6	60.8	56.4
	3.0	58.4	61.2	55.6
	0.5	57.8	59.8	55.7
	0.33	57.3	58.8	55.8
	0.1	53.9	54.6	53.1
Batch size	256	59.3	61.6	57.1
	1024	58.6	60.0	57.3
	512	58.6	60.8	56.4
	128	58.1	61.1	55.1
Initialization	FS	62.7	64.4	60.9
	Rand	58.6	60.8	56.4
	SS	57.9	58.9	56.8
Epochs	300	58.6	60.8	56.4
	100	58.4	60.7	56.1
	200	58.3	60.3	56.4
	50	55.5	58.0	53.0

Ablation	Setting				Selection metric	Phase (F1)	Tool (mAP)
	MC	C	G	S			
Augmentations	2	✓	✓	✗	60.0	63.5	56.5
	2	✓	✓	✓	59.6	63.2	55.9
	4	✓	✓	✓	59.1	61.7	56.5
	4	✓	✓	✗	58.9	61.1	56.8
	8	✓	✓	✗	58.6	60.8	56.4
	8	✓	✓	✓	54.7	56.4	53.1
	2	✗	✓	✗	53.7	55.4	52.0
	2	✗	✓	✓	53.3	54.6	51.9
					...		
	8	✓	✗	✓	45.5	47.5	43.6
	4	✗	✗	✓	41.2	42.2	40.2
	2	✗	✗	✓	40.2	41.1	39.4
	8	✗	✗	✓	37.3	37.8	36.8
	4	✗	✗	✗	37.3	36.9	37.6
	2	✗	✗	✗	37.0	37.2	36.8
	8	✗	✗	✗	36.8	35.8	37.7
	8	✗	✓	✓	33.1	31.4	34.8

with higher sampling rates (Fig. 11). This is an important finding that may lend useful intuition to researchers applying SSL to domains with similar motion characteristics on how best to allocate computational resources, when training these intensive methods comes with a sizeable financial and environmental cost. To note, for a fair comparison, we perform this experiment here assuming an equal distribution of computational resources, i.e. we evaluated the models after performing self-supervised pretraining for the same number of iterations for each frame rate. This implies that the 1 fps experiments were trained for ~ 5 times as many epochs as the 5 fps experiments.

Initialization. In general computer vision, the common practice for SSL experimentation is to train models to learn self-supervised representations entirely from scratch (i.e. random weights) before using these representations to attempt to replicate fully supervised performance - for image recognition on Imagenet, as a prominent example. Weights obtained in this manner are then intended to serve as initialization for downstream tasks. However, in surgical computer vision, Imagenet fully supervised weights are considered as a readily available resource: the practice of using them to initialize models is tacitly recognized as standard by the community. The choice of initialization is therefore not trivial, with 3 options available before starting SSL training on surgical data:

1. "Rand.": randomly initializing weights
2. "SS": initializing weights with self-supervised pretraining on ImageNet
3. "FS": initializing weights with fully supervised pretraining on ImageNet

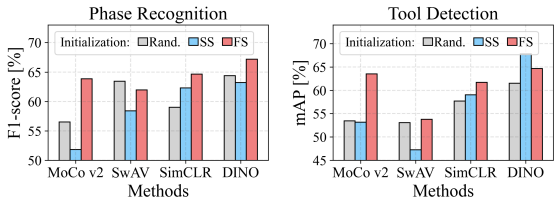

Fig. 12. Performance of each method on Cholec80 varying the network initialization strategies before performing self-supervised pretraining: random initialization (*Rand.*), ImageNet self-supervised (*SS*), ImageNet fully-supervised (*FS*). Results were obtained using linear evaluation on the validation set. Left: F_1-score for phase recognition. Right: mAP for tool presence detection.

Across all SSL methods (Fig. 12), models initialized with "FS" significantly outrank models with "Rand." or "SS." initialization; most noticeably with MoCo v2 (up to +12% phase recognition F_1, +11% tool detection mAP compared to the other two). Results between "Rand." and "SS." do not clearly favor one over the other. This is obviously a major difference from general computer vision, which expects models initialized from scratch to improve on any downstream task through SSL training. One explanation for this discrepancy could be the set of invariances learned in the natural domain, which may not apply to surgical images.

Hyperparameter study conclusion. This study provides a detailed view of each SSL method's reaction to changes in parametrization when operating in the surgical domain, exposing noteworthy differences with the natural domain - regarding augmentations, batch size and initialization most prominently. However, when considering all four SSL methods and both tasks simultaneously, global trends can be difficult to clearly point out. To achieve this in a quantitative and principled manner, we define a selection metric, defined as the average

of all phase recognition F_1 scores and tool presence detection mAPs across all methods for a given setting. Using this, we are able to rank the values of a given hyperparameter by overall performance across downstream tasks, and then retain the best. This forms a global set of **recommended settings** (Table 1) for SSL in the surgical domain.

In Table 2, we present the results ranked according to this selection metric for each ablation to facilitate the analysis of invariant trends for methods and tasks. For each hyperparameter, we summarize the trends in brief below:

- Sampling rate: We observe only a marginal utility of increasing the sampling rate beyond a certain point, with the selection metric saturating past 0.33 fps.
- Batch size: The results show that for the considered tasks and dataset, SSL method performance is mostly robust to variations of batch size. Varying the batch size between 128-1024 results in a maximum variation of 1.1% F1 and 2% mAP on average across methods for phase recognition and tool presence detection, respectively.
- Initialization: Initialization before self-supervised representation learning proves to be a critical hyperparameter with significant and consistent gains in performance across both methods and tasks. Initializing with Imagenet fully supervised ("FS") weights proves to be the optimal setting amongst the considered initializations.
- Epochs: For both considered tasks, we see significant gains in performance up to 100 epochs after which it plateaus, with an average variation of 0.4% F1 and 0.4% mAP between 100 and 300 epochs.
- Augmentations: Interestingly, we observe largely consistent trends for different augmentation settings for both tasks. Color and geometric augmentations feature consistently in top-performing augmentation settings. On average across methods, the addition of multiple low-resolution views and strong color augmentations has a less clear impact on performance.

4.3. Data supply study

The recommended choice of hyperparameters mentioned above provides, on average, close to optimal conditions for observing our panel of SSL methods in practical use cases, with varying quantities of labeled or unlabeled data. Our proposed usage of SSL is defined as follows: self-supervised training is performed in the surgical domain before finetuning for surgical downstream tasks.

Labeled data supply. In this section of the data supply study, self-supervised training is first performed on the entire training set of Cholec80 with the recommended hyperparameters. Surgical downstream task finetuning is then applied using variable amounts of labeled data: 40 videos (100% of the training set), or in semi-supervision with 10 videos (25%) or 5 videos (12.5%); for these last two settings, the portions of the training set are drawn following a stratified random sampling approach (see Sec. 4.1). Results for these experiments are reported in Tables 3 (phase recognition on single frames), 4 (phase recognition on videos with a temporal model), and 5 (tool presence detection). We compare our proposed usage of SSL ("ours") on Cholec80 using the recommended hyperparameters (Table 1) with the mode of operation borrowed from general computer vision ("base") - i.e. finetuning directly from weights pretrained with SSL on Imagenet. The bottom row in each table ("No SSL") provides an additional point of comparison, where we finetune models initialized with fully supervised Imagenet weights without any SSL.

In most low-label settings (10, 5 videos), adding any of the 4 SSL methods systematically improves performance on both surgical tasks, compared to direct finetuning from supervised Imagenet weights without SSL. This improvement reaches up to 6.1% (5 videos, MoCo v2) for single-frame phase recognition, 6% (5 videos, SwAV) for temporal phase recognition, and 14.7% for tool presence detection (5 videos, MoCo v2). Gains are consistently observed, especially in low-label settings where standard deviation across splits mostly stays underneath 3% (32 out of 48 table entries). 100% label availability tends to saturate performance on downstream tasks, leaving little room for improvement from SSL; still, results are on par with those obtained without SSL for both tool presence detection (mostly $<$ 1% difference) and phase recognition, with the largest deficit (-1.2%) recorded for SwAV on single frames. Out of the four SSL methods presented here, MoCo v2 seems to yield better results, 5 times achieving the best performance for a given number of labeled videos.

Most importantly, these results challenge the generalizability of general computer vision SSL. As demonstrated in Oord et al. (2018); He et al. (2020); Chen et al. (2020c); Caron et al. (2021), self-supervised pretraining on natural images enhances downstream task performance in the natural image domain; however, these gains may not carry over to more complex and more specific domains. Indeed, when pretrained on Imagenet, rarely do any of the SSL methods featured here improve performance on surgical downstream tasks, compared to the "No SSL" baseline (only 7 out of 36 times). For phase recognition, this usage of SSL can cause F_1 score to drop by up to 1.9%, while for tool presence detection, the degradation reaches up to 11.2% mAP. Overall, our proposed use of SSL outperforms the "base" usage by up to 6.2% on single-frame phase recognition, 7.4% on temporal phase recognition, and 20.4% on tool presence detection.

Finally, we add an external comparison in Table 6 with preexisting semi-supervised studies in surgical computer vision, based on results presented by Shi et al. (2021) for semi-supervised phase recognition on Cholec80, and using the same split and metric definition. As expected, selected SSL methods applied to single-frame models are often outranked by other approaches, by up to 16.6% (DINO vs SurgSSL, 10 videos); however the external methods, we compare against, use temporal modeling, which gives them a strong advantage. For a fairer comparison, we examine models trained with our selected SSL methods used in conjunction with a temporal

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

Table 3. Effect of our proposed SSL pretraining in the surgical domain ("Ours") on surgical phase recognition performance from single frames. "Base" refers to self-supervised pretraining on Imagenet only. "No SSL" refers to fully supervised pretraining on Imagenet only. Bold indicates the best performance for a given number of labeled videos.

Surgical phase recognition F_1 - single frame						
Labels	40 videos		10 videos		5 videos	
	Base	**Ours**	Base	**Ours**	Base	**Ours**
DINO	71.6	71.1	60.6 ± 0.6	62.2 ± 0.9	51.4 ± 5.1	56.3 ± 4.8
MoCo v2	70.3	71.3	58.5 ± 0.6	**64.4 ± 1.7**	52.1 ± 4.5	**58.1 ± 5.3**
SimCLR	70.3	**71.8**	58.9 ± 2.4	63.5 ± 1.1	51.3 ± 3.9	57.2 ± 5.0
SwAV	70.2	70.3	58.8 ± 0.9	62.2 ± 1.9	50.9 ± 4.5	57.1 ± 3.7
No SSL	71.5		60.4 ± 0.4		52.0 ± 6.5	

Table 4. Effect of our proposed SSL pretraining in the surgical domain ("Ours") on surgical phase recognition performance from videos when finetuning a temporal model (TCN - Czempiel et al. (2020)) on top of the backbones described in Table 3. Bold indicates the best performance for a given amount of labeled videos.

Surgical phase recognition F_1 - temporal						
Labels	40 videos		10 videos		5 videos	
	Base	**Ours**	Base	**Ours**	Base	**Ours**
DINO	81.5	**81.6**	71.3 ± 0.6	70.4 ± 0.4	61.1 ± 9.0	65.0 ± 5.4
MoCo v2	79.5	79.6	69.1 ± 1.8	**74.1 ± 0.4**	63.4 ± 4.3	66.1 ± 4.2
SimCLR	78.8	81.1	69.2 ± 2.4	72.5 ± 0.4	63.6 ± 3.9	66.6 ± 2.4
SwAV	78.4	79.5	68.7 ± 0.5	71.4 ± 0.7	60.9 ± 7.0	**68.3 ± 1.3**
No SSL	80.3		70.1 ± 0.2		62.3 ± 7.4	

Table 5. Effect of our proposed SSL pretraining in the surgical domain ("Ours") on surgical tool presence detection performance. Bold indicates the best performance for a given amount of labeled videos.

Surgical tool presence detection mAP						
Labels	40 videos		10 videos		5 videos	
	Base	**Ours**	Base	**Ours**	Base	**Ours**
DINO	92.1	93.2	70.1 ± 2.7	81.2 ± 1.4	50.6 ± 1.6	68.7 ± 2.3
MoCo v2	92.9	93.5	70.4 ± 1.3	**85.7 ± 1.1**	56.5 ± 3.3	**74.7 ± 1.8**
SimCLR	90.4	93.1	66.7 ± 0.1	83.0 ± 0.9	49.3 ± 1.4	69.7 ± 3.0
SwAV	92.5	92.8	70.5 ± 1.5	79.1 ± 1.7	52.5 ± 1.8	63.0 ± 0.7
No SSL	**93.6**		77.9 ± 0.8		60.0 ± 2.3	

Table 6. External comparison with Shi et al. (2021) for semi-supervised surgical phase recognition. Bold indicates the best performance for a given amount of labeled videos used for finetuning.

External comparison - surgical phase recognition F_1					
Labels			40 videos	10 videos	5 videos
External	**NL-RCNet**		82.1	73.5	67.3
quoted from Shi et al. (2021)	**NL-RCNet+**		84.4	-	-
	CNN-BiLSTM-CRF		-	75.3	70.9
	MT		-	77.3	71.0
	SurgSSL		-	80.6	78.6
Selected SSL methods	**DINO**	single frame	77.6	64.0	65.4
metric and split from Shi et al. (2021)		temporal	91.8	81.1	76.9
	MoCo v2	single frame	81.7	72.6	69.3
		temporal	91.3	82.5	**81.4**
	SimCLR	single frame	84.5	73.8	67.0
		temporal	**93.6**	**85.0**	80.0
	SwAV	single frame	86.1	67.1	69.5
		temporal	91.0	79.8	80.7
Baselines	**No SSL**	single frame	81.0	65.6	60.8
metric and split from Shi et al. (2021)		temporal	87.4	81.5	78.4

model (TCN): in these situations, they surpass preexisting semi-supervised approaches by a substantial amount - up to 14.1%. Top F_1 scores are achieved by SimCLR (93.6%, labels on 40 videos - 85.0%, labels on 10 videos) and MoCo v2 (81.4%, labels on 5 videos). To note, the architecture we use is fairly simple (CNN - TCN) compared to the more refined designs featured in the external methods; therefore our performance gains derive from the SSL methodology itself, and could further increase with more advanced architectures. These observations strongly confirm the high value of bringing SSL innovations from general computer vision to the surgical domain.

Unlabeled data supply. Our main experiments examined the performance of SSL in the surgical domain with a fixed quantity of unlabeled data for self-supervised pretraining; in this complementary set of experiments, we observe how SSL reacts when the quantity of unlabeled videos varies. This part of the study is conducted with MoCo v2 exclusively. Overall, our results (Fig. 13 and 14) confirm a valuable benefit of SSL: for the most part, expanding unlabeled data - which is far easier than generating additional annotations - leads to increased performance in downstream surgical tasks. Particularly when few labeled instances are available, we see extremely pronounced improvements brought about by introducing SSL. For example, when only 5 labeled videos are available, self-supervised pretraining on just 10 unlabeled videos adds 4.2% F_1 for phase recognition and ~14.2% mAP for tool presence detection. These results further reinforce the practicality of utilizing these SSL methods in surgical applications, where working with small datasets is often the norm rather than the exception. We observe, however, two main limitations.

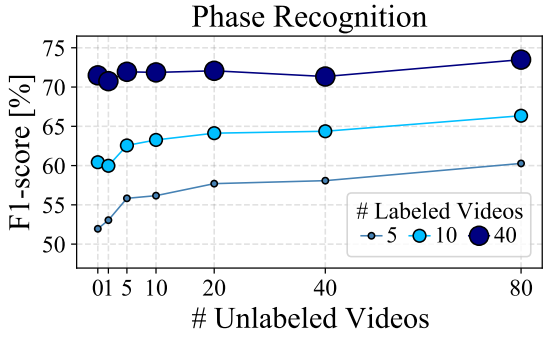

Fig. 13. Single-frame phase recognition performance of MoCo v2 w.r.t. the number of unlabeled videos used for self-supervised pretraining, with finetuning on 5, 10, and 40 labeled videos.

The first is a *saturation* phenomenon, apparent after 10 unlabeled videos; while going from 1 unlabeled video to 10 clearly improves feature quality (phase recognition, finetuning on 5 labeled: +3.1% F_1; tool presence detection, finetuning on 5 labeled: +9.3% mAP), results for 10 and up carry more ambiguity, with large differences depending on the task. While phase recognition performance slows down but still

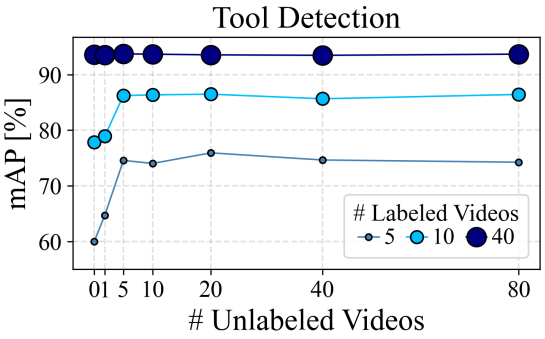

Fig. 14. Tool presence detection performance of MoCo v2 w.r.t. the number of unlabeled videos used for self-supervised pretraining, with finetuning on 5, 10, and 40 labeled videos.

increases by a noticeable amount (e.g. finetuning on 5 labeled, +4.1% F_1 from 10 to 80), tool presence detection completely halts.

The second is *dilution* by labeled data: using larger amounts of annotated videos for finetuning pushes downstream performance closer to its limits, which tends to equalize the effect of adding unannotated videos. For example, for phase recognition from 1 to 80 unlabeled videos, F_1 score increases by 7.2% with 5 labeled but only by 2.7% with 40 labeled. Dilution is much stronger for tool presence detection: from 1 to 80 unlabeled, the total mAP increase with 5 labeled is 9.5%, while no gain is perceivable at all with 40 labeled.

As evidenced by these observations, the performance growth brought by SSL can slow down as the unlabeled data supply increases, depending on the amount of annotated data available as well as the nature of the task. Tool labels are tied to distinct pieces of visual evidence in the image; their influence on the model's final performance is therefore extremely high, compared to unlabeled videos used in self-supervision. In contrast, phase labels tend to accompany more ambiguous visual cues, which would explain why the advantage of using SSL is much more apparent for surgical phase recognition: a model pretrained with 80 unlabeled videos and finetuned on only 5 labeled videos reaches 60.3% F_1, which is about the same as a model pretrained with 1 unlabeled but finetuned on 10 labeled. Saturation for phase recognition is also much softer than for tool presence detection, suggesting performance can increase even further with more than 80 videos.

4.4. Generalization study

Using the same recommended hyperparameters established in Section 4.2, we conduct experiments using MoCo v2 on the collection of datasets presented in Section 3.5. Results are presented in Table 7 demonstrating how SSL representations could be adapted for data from other sources and for other vision-based tasks.

HeiChole Experiments. In this first experiment series of the generalization study, we utilize the HeiChole Benchmark

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

Exp #	Dataset - architecture			Generalization Experiments					
				Labeled videos		Labeled videos		Labeled videos	
	SSL Dataset	Task	Metric	No SSL	**MoCo v2**	No SSL	**MoCo v2**	No SSL	**MoCo v2**
	HeiChole - TCN head			24 videos		4 videos		2 videos	
1	Cholec80	Phase	F_1	58.6	**64.7**	41.7 ± 4.7	**51.1 ± 3.3**	27.6 ± 6.0	**39.0 ± 1.2**
	HeiChole - linear head								
2	Cholec80	Tool	mAP	62.5	**66.9**	36.7 ± 2.9	**43.7 ± 0.4**	25.1 ± 6.1	**30.3 ± 2.3**
	CATARACTS - TCN head			25 videos		6 videos		3 videos	
3	Cholec80	Phase	F_1	**75.2**	74.5	**65.7 ± 5.5**	65.0 ± 5.6	**52.8 ± 4.7**	50.7 ± 1.0
4	CATARACTS	Phase	F_1	75.2	**77.2**	65.7 ± 5.5	**66.5 ± 3.8**	52.8 ± 4.7	**56.2 ± 5.5**
	CATARACTS - linear head								
5	Cholec80	Tool	mAP	**56.1**	47.6	**37.7 ± 1.4**	29.2 ± 2.2	**26.9 ± 1.6**	19.0 ± 0.4
6	CATARACTS	Tool	mAP	56.1	**57.3**	37.7 ± 1.4	**40.8 ± 0.5**	26.9 ± 1.6	**31.2 ± 4.2**
	CholecT50 - linear head			40 videos		10 videos		5 videos	
7	Cholec80	Action	mAP	19.4	**26.7**	14.4 ± 0.2	**20.7 ± 0.2**	11.2 ± 1.4	**15.9 ± 0.8**
	CholecT50 - RDV head								
8	Cholec80	Action	mAP	31.4	**35.7**	22.3 ± 1.8	**25.5 ± 0.8**	14.9 ± 0.9	**18.3 ± 1.2**
	Endoscapes - DeepLabv3+ head			120 videos		30 videos		15 videos	
9	Cholec80	Segmentation	F_1	**73.2**	**73.2**	63.6 ± 1.0	**64.3 ± 1.0**	58.1 ± 1.2	**59.3 ± 1.7**
	CaDIS 8 classes - DeepLabv3+ head			19 videos		4 videos		2 videos	
10	Cholec80	Segmentation	F_1	86.9	**87.1**	79.6 ± 1.6	**82.5 ± 1.2**	79.5 ± 1.6	**81.4 ± 1.2**
11	CaDIS	Segmentation	F_1	**86.9**	**86.9**	79.6 ± 1.6	**83.2 ± 0.8**	79.5 ± 1.6	**81.3 ± 0.8**
	CaDIS 25 classes - DeepLabv3+ head								
12	Cholec80	Segmentation	F_1	**71.8**	70.5	61.2 ± 1.9	**62.4 ± 2.9**	55.5 ± 5.8	**57.3 ± 6.7**
13	CaDIS	Segmentation	F_1	**71.8**	71.7	61.2 ± 1.9	**61.6 ± 2.8**	55.5 ± 5.8	**56.5 ± 5.7**

Table 7. Results on additional data & tasks; finetuning directly from ImageNet pretrained weights (No SSL) vs finetuning after MoCo V2 pretraining. In each experiment, we state the model architecture placed after the ResNet50 backbone, the SSL dataset used to pretrain the backbone, and the task and metric under consideration. For each dataset, we also conduct experiments with 3 subsets of labeled videos used for training.

for surgical workflow analysis. Similar to Cholec80, this HeiChole dataset comprises videos for surgical phase recognition and tool presence detection for laparoscopic cholecystectomy. This serves as an ideal benchmark to evaluate how self-supervised representations learned from similar data (same procedure) could be used to boost performance for vision-based tasks on independently sourced datasets with potentially varying surgical workflows, acquisition methods, instrumentation, etc. Indeed, experiments 1 and 2 in Table 7 reveal significant boosts in performance when initializing from models pretrained on Cholec80 (using SSL) at all considered proportions of labeled data. Most notably, using only 2 labeled videos, we observe boosts in performance of 11.4% for phase recognition and 5.2% for tool presence detection. Based on the official leaderboard of the HeiChole challenge, presented in Table 9, this would have positioned our method in 1^{st} place for the tool presence detection task and 4^{th} for surgical phase recognition using only a simple model architecture. These results strongly exemplify the impact that SSL methods, such as the ones investigated in this article, could have on learning from small datasets and datasets with underrepresented characteristics, problems endemic to surgical data science (Maier-Hein et al., 2022).

CATARACTS Experiments. Similar to the HeiChole benchmark, the CATARACTS dataset introduces two similar tasks for surgical workflow recognition but with two notable differences: (1) The CATARACTS datasets depict scenes from cataract surgery procedures with a strikingly different appearance and workflow from laparoscopic cholecystectomy (2) The temporal task introduced with this dataset is surgical step recognition, which normally refers to finer temporal segments than surgical phases (Mascagni et al., 2022). This series of experiments reveals two important findings. Firstly, unlike the HeiChole experiments, models pretrained on Cholec80 (Table 7, experiments 3 and 5) consistently perform worse than models initialized from Imagenet ("No SSL"). This may be attributed to the significantly distinct and specific visual appearance of Cholec80 scenes serving as a confounding factor when learning representations. However, we do note that when initializing from SSL weights learned on CATARACTS, we see consistent boosts of $\sim 1-4\%$ compared to Imagenet initializations across both the downstream tasks. This provides an indication that the SSL setup presented in this work could be adapted to other surgical datasets without further hyperparameter tuning for the pretraining stage.

CholecT50 Experiments. In this series of experiments, we aim to illustrate how self-supervised representations could also help in more difficult workflow tasks like action recognition. To this end, we evaluate performance on CholecT50, a large dataset of surgical actions annotated on videos sourced from the same hospital as Cholec80. Note that the action triplet recognition task on CholecT50 is performed twice (Table 7, experiments 7 and 8): once using a simple linear head, then a second time with Nwoye et al. (2022b)'s Rendezvous

CholecTriplet 2021 challenge leaderboard	
Rank	Triplet recognition mAP
1^{st}	38.1
2^{nd}	35.8
MoCo V2 - RDV head	**35.7**
3^{rd}	32.9
4^{th} *(RDV baseline)*	32.7

Table 8. Comparison of MoCo v2 pretraining against the official top 4 entries in the 2021 CholecTriplet challenge.

HeiChole Benchmark			
Rank	Phase (F_1)	Rank	Tool (F_1)
1	68.8	**MoCo V2**	**66.9**
2	65.4	1	63.8
3	65.0	2	63.0
MoCo V2	**64.7**	3	58.2
4	63.6	4	50.1

Table 9. Comparison of MoCo v2 pretraining against the official top 4 entries for the phase and tool tasks in the HeiCholec Benchmark (EndoVis challenge 2019).

(RDV) head. In both settings, we observe consistent and marked boosts in performance at all proportions of labeled data demonstrating the utility of these methods across model design choices. Most impressively, utilizing a previously published architecture (Nwoye et al., 2022b) with a generic initialization of features would have placed 3^{rd} (Table 8) in the CholecTriplet 2021 challenge (Nwoye et al., 2022a), further illustrating the value that SSL could bring to the surgical data science community.

Segmentation Experiments. Here, we aim to explore how self-supervised representations also have utility for tasks requiring more spatial reasoning than frame-level classification. To this end, we use two surgical semantic segmentation datasets: Endoscapes, consisting of laparoscopic cholecystectomy videos sourced from the same hospital as Cholec80, CaDIS 8 classes and CaDIS 25 classes, containing cataract surgery videos. Consistently, across all three segmentation tasks and labeled data settings, we observe trends consistent with previous findings: pretraining models using SSL deliver boosts in performance. However, the performance boosts are generally less pronounced than the other considered image recognition tasks. This may be because the considered SSL methods define the learning problem by considering global-level features from the complete image. However, semantic segmentation requires more dense spatial reasoning. More specific architectures choices (Caron et al., 2021) or SSL methods (Wang et al., 2021; Xie et al., 2022) could further improve downstream segmentation performance.

5. Conclusion

Despite major progress in the field of self-supervised representation learning over the last several years, its adoption into label-scarce fields like surgery, where it could perhaps have the most significant impact, has been slow. This could be due to the demonstrably heavy reliance on hyperparameter choices that SSL methods demand. In this paper, we conduct an extensive benchmark study to methodically identify effective hyperparameter settings for the task of surgical phase recognition and tool presence detection on the Cholec80 dataset. From this strong foundation, we deployed SSL on a highly diverse array of surgical datasets, obtaining solid results that support its use for many surgical vision tasks.

Requiring over 7000 GPU hours, the hyperparameter study demonstrates that this exploration is pivotal to the practical utility of SSL in settings such as semi-supervised learning. For example, initializing the base architecture using Imagenet weights before SSL pretraining critically provided consistent, marked boosts in performance over all other initializations. While random initialization before performing self-supervised representation learning is the standard practice in other large studies, perhaps because of the relative size of the considered datasets, this example highlights the need for principled, adaptable methods to identify optimal settings for other domains. Additionally, domain characteristics could indicate the most significant parameters to prioritize for searches. For instance, in our experiments, relatively slow motion patterns may explain why sampling frames at higher rates for representation learning provides little to no improvement beyond a certain point.

In the data supply study, SSL pretraining shows promising boosts in performance for all methods, particularly in label-scarce scenarios for both phase recognition and tool presence detection. Interestingly, these methods even outperform state-of-the-art methods for semi-supervised phase recognition using only generic representational features. These results are strongly indicative of the value of targeting surgical applications using these SSL methods, which, within certain limits, can be enhanced by simply incorporating additional unannotated data.

The generalization study displays the full strength of SSL, with strong results across many surgical contexts; again with generic features obtained without labels. Excellent robustness is demonstrated when switching to a different clinical center or to another task - even the most fine-grained. Results obtained on cataract surgery with hyperparameters conserved from cholecystectomy are highly encouraging for even more radical generalizations of SSL. Further, experimental validation on public challenges, a popular format to introduce and benchmark new datasets, revealed that even simple model architectures with "generic" SSL-based initializations achieve more than competitive results compared to significantly more sophisticated design choices. This is despite a recent survey (Eisenmann et al., 2022) concluding that a median of 80 working hours and 267 GPU hours were dedicated in such challenges to model development and training, respectively. Overall, this section of the study presents a strong exemplification of the value and impact that SSL methods, such as the ones described in this work, could have on supporting ongoing efforts in surgical data science, where small datasets with underrepresented characteristics and expensive annotations are a common occurrence.

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

Out of the many possibilities opened up by this study, two stand out as highly promising directions for future work: the first one is federated learning (McMahan et al., 2017), where SSL can play a major role by learning robust features from data scattered across multiple clinical centers (Kassem et al., 2022). Another natural progression from this work is to apply these findings to recent work in spatio-temporal representation learning and adapt them to the unique characteristics of surgical videos.

Finally, we note that only a select subset of trends were presented for analysis in this work due to many being results aggregated across methods, splits, or other experimental settings for brevity. With around 500 experiments run over 9000 GPU hours, we will disclose complete results for the experiments conducted in this work, in order to facilitate future research on SSL in surgery. The code, along with results and checkpoints, is available at https://github.com/CAMMA-public/SelfSupSurg.

Acknowledgements This work was partially supported by French state funds managed by the ANR under references ANR-20-CHIA-0029-01 (National AI Chair AI4ORSafety), ANR-10-IAHU-02 (IHU Strasbourg) and ANR-16-CE33-0009 (DeepSurg). This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813782 - project ATLAS. This work was supported by a Ph.D. fellowship from Intuitive Surgical. It was granted access to the HPC resources of IDRIS under the allocations 2021-AD011011638R1, 2021-AD011011638R2, 2021-AD011012715, 2021-AD011012832, 2021-AD011011507R1, and 2021-AD011011640R1. For evaluation on the HeiChole dataset, we thank Dr. Sebastian Bodenstedt for the timely support.

References

Ahsan, U., Madhok, R., Essa, I.A., 2019. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition, in: IEEE Winter Conference on Applications of Computer Vision, WACV 2019, IEEE. pp. 179–189. URL: https://doi.org/10.1109/WACV.2019.00025, doi:10.1109/WACV.2019.00025.

Al Hajj, H., Lamard, M., Conze, P.H., Cochener, B., Quellec, G., 2018. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. Medical image analysis 47, 203–218.

Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., et al., 2019. Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical image analysis 52, 24–41.

Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B., Padoy, N., 2021. Temporally constrained neural networks (TCNN): A framework for semi-supervised video semantic segmentation. CoRR abs/2112.13815. URL: https://arxiv.org/abs/2112.13815, arXiv:2112.13815.

Asano, Y.M., Rupprecht, C., Vedaldi, A., 2019. Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371 .

Bachman, P., Hjelm, R.D., Buchwalter, W., 2019. Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910 .

Bao, H., Dong, L., Piao, S., Wei, F., 2022. BEit: BERT pre-training of image transformers, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=p-BhZSz59o4.

Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T., 2020. Speednet: Learning the speediness in videos, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE. pp. 9919–9928. doi:10.1109/CVPR42600.2020.00994.

Blum, T., Feußner, H., Navab, N., 2010. Modeling and segmentation of surgical workflow from laparoscopic video, in: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2010, 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part III, Springer. pp. 400–407. URL: https://doi.org/10.1007/978-3-642-15711-0_50, doi:10.1007/978-3-642-15711-0_50.

Bodenstedt, S., Wagner, M., Katic, D., Mietkowski, P., Mayer, B., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S., 2017. Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis. arXiv .

Boutillon, A., Conze, P., Pons, C., Burdin, V., Borotikar, B., 2021. Multi-task, multi-domain deep segmentation with shared representations and contrastive regularization for sparse pediatric datasets, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part I, Springer. pp. 239–249. URL: https://doi.org/10.1007/978-3-030-87193-2_23, doi:10.1007/978-3-030-87193-2_23.

Caron, M., Bojanowski, P., Joulin, A., Douze, M., 2018. Deep clustering for unsupervised learning of visual features, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 132–149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 .

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 .

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society. pp. 4724–4733. URL: https://doi.org/10.1109/CVPR.2017.502, doi:10.1109/CVPR.2017.502.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2020a. Generative pretraining from pixels, in: International Conference on Machine Learning, PMLR. pp. 1691–1703.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, X., Fan, H., Girshick, R., He, K., 2020c. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X., 2021. USCL: pretraining deep ultrasound image diagnosis model through video contrastive representation learning, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VIII, Springer. pp. 627–637. URL: https://doi.org/10.1007/978-3-030-87237-3_60, doi:10.1007/978-3-030-87237-3_60.

da Costa Rocha, C., Padoy, N., Rosa, B., 2019. Self-supervised surgical tool segmentation using kinematic information, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE. pp. 8720–8726.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q., 2020a. Randaugment: Practical automated data augmentation with a reduced search space, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020b. Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703.

Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transport, in: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 2292–2300. URL: https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks, in: MICCAI.

Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N., 2021. Opera: Attention-regularized transformers for surgical phase recognition, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part IV, Springer. pp. 604–614. URL: https://doi.org/10.1007/978-3-030-87202-1_58, doi:10.1007/978-3-030-87202-1_58.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., Jannin, P., 2016. Automatic data-driven real-time segmentation and recognition of surgical workflow. Int. J. Comput. Assist. Radiol. Surg. 11, 1081–1089. URL: https://doi.org/10.1007/s11548-016-1371-x, doi:10.1007/s11548-016-1371-x.

Diba, A., Sharma, V., Gool, L.V., Stiefelhagen, R., 2019. Dynamonet: Dynamic action and motion network, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE. pp. 6191–6200. URL: https://doi.org/10.1109/ICCV.2019.00629, doi:10.1109/ICCV.2019.00629.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, pp. 1422–1430.

Dong, N., Voiculescu, I., 2021. Federated contrastive learning for decentralized unlabeled medical images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III, Springer. pp. 378–387. URL: https://doi.org/10.1007/978-3-030-87199-4_36, doi:10.1007/978-3-030-87199-4_36.

Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Peco, N.Y., 2021. Perceptual codebook for bert pre-training of vision transformers. arXiv preprint arXiv:2111.12710 1, 7.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net. URL: https://openreview.net/forum?id=YicbFdNTTy.

Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T., 2014a. Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing systems 27, 766–774.

Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T., 2014b. Discriminative unsupervised feature learning with convolutional neural networks, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, MIT Press, Cambridge, MA, USA. p. 766–774.

Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., McDonald, C., Piguet, C.M., Phillips, M.L., Eyler, L., Duchesnay, E., 2021. Contrastive learning with continuous proxy meta-data for 3d MRI classification, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 58–68. URL: https://doi.org/10.1007/978-3-030-87196-3_6, doi:10.1007/978-3-030-87196-3_6.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2650–2658. doi:10.1109/ICCV.2015.304.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Godau, P., Cheplygina, V., Kozubek, M., Ali, S., Gupta, A., Kybic, J., Noble, A., de Solórzano, C.O., Pachade, S., Petitjean, C., Sage, D., Wei, D., Wilden, E., Alapatt, D., Andrearczyk, V., Baid, U., Bakas, S., Balu, N., Bano, S., Bawa, V.S., Bernal, J., Bodenstedt, S., Casella, A., Choi, J., Commowick, O., Daum, M., Depeursinge, A., Dorent, R., Egger, J., Eichhorn, H., Engelhardt, S., Ganz, M., Girard, G., Hansen, L., Heinrich, M., Heller, N., Hering, A., Huaulmé, A., Kim, H., Landman, B., Li, H.B., Li, J., Ma, J., Martel, A., Martín-Isla, C., Menze, B., Nwoye, C.I., Oreiller, V., Padoy, N., Pati, S., Payette, K., Sudre, C., van Wijnen, K., Vardazaryan, A., Vercauteren, T., Wagner, M., Wang, C., Yap, M.H., Yu, Z., Yuan, C., Zenk, M., Zia, A., Zimmerer, D., Bao, R., Choi, C., Cohen, A., Dzyubachyk, O., Galdran, A., Gan, T., Guo, T., Gupta, P., Haithami, M., Ho, E., Jang, I., Li, Z., Luo, Z., Lux, F., Makrogiannis, S., Müller, D., Oh, Y.t., Pang, S., Pape, C., Polat, G., Reed, C.R., Ryu, K., Scherr, T., Thambawita, V., Wang, H., Wang, X., Xu, K., Yeh, H., Yeo, D., Yuan, Y., Zeng, Y., Zhao, X., Abbing, J., Adam, J., Adluru, N., Agethen, N., Ahmed, S., Khalil, Y.A., Alenyà, M., Alhoniemi, E., An, C., Anwar, T., Arega, T.W., Avisdris, N., Aydogan, D.B., Bai, Y., Calisto, M.B., Basaran, B.D., Beetz, M., Bian, C., Bian, H., Blansit, K., Bloch, L., Bohnsack, R., Bosticardo, S., Breen, J., Brudfors, M., Brüngel, R., Cabezas, M., Cacciola, A., Chen, Z., Chen, Y., Chen, D.T., Cho, M., Choi, M.K., Xie, C.X.C., Cobzas, D., Cohen-Adad, J., Acero, J.C., Das, S.K., de Oliveira, M., Deng, H., Dong, G., Doorenbos, L., Efird, C., Fan, D., Serj, M.F., Fenneteau, A., Fidon, L., Filipiak, P., Finzel, R., Freitas, N.R., Friedrich, C.M., Fulton, M., Gaida, F., Galati, F., Galazis, C., Gan, C.H., Gao, Z., Gao, S., Gazda, M., Gerats, B., Getty, N., Gibicar, A., Gifford, R., Gohil, S., Grammatikopoulou, M., Grzech, D., Güley, O., Günnemann, T., Guo, C., Guy, S., Ha, H., Han, L., Han, I.S., Hatamizadeh, A., He, T., Heo, J., Hitziger, S., Hong, S., Hong, S., Huang, R., Huang, Z., Huellebrand, M., Huschauer, S., Hussain, M., Inubushi, T., Polat, E.I., Jafaritadi, M., Jeong, S., Jian, B., Jiang, Y., Jiang, Z., Jin, Y., Joshi, S., Kadkhodamohammadi, A., Kamraoui, R.A., Kang, I., Kang, J., Karimi, D., Khademi, A., Khan, M.I., Khan, S.A., Khantwal, R., Kim, K.J., Kline, T., Kondo, S., Kontio, E., Krenzer, A., Kroviakov, A., Kuijf, H., Kumar, S., La Rosa, F., Lad, A., Lee, D., Lee, M., Lena, C., Li, H., Li, L., Li, X., Liao, F., Liao, K., Oliveira, A.L., Lin, C., Lin, S., Linardos, A., Linguraru, M.G., Liu, H., Liu, T., Liu, D., Liu, Y., Lourenço-Silva, J., Lu, J., Lu, J., Luengo, I., Lund, C.B., Luu, H.M., Lv, Y., Lv, Y., Macar, U., Maechler, L., L., S.M., Marshall, K., Mazher, M., McKinley, R., Medela, A., Meissen, F., Meng, M., Miller, D., Mirjahanmardi, S.H., Mishra, A., Mitha, S., Mohyud Din, H., Mok, T.C.W., Murugesan, G.K., Karthik, E.N., Nalawade, S., Nalepa, J., Naser, M., Nateghi, R., Naveed, H., Nguyen, Q.M., Quoc, C.N., Nichyporuk, B., Oliveira, B., Owen, D., Pal, J.B., Pan, J., Pan, W., Pang, W., Park, B., Pawar, V., Pawar, K., Peven, M., Philipp, L., Pieciak, T., Plotka, S., Plutat, M., Pourakpour, F., Preložnik, D., Punithakumar, K., Qayyum, A., Queirós, S., Rahmim, A., Razavi, S., Ren, J., Rezaei, M., Rico, J.A., Rieu, Z., Rink, M., Roth, J., Ruiz-Gonzalez, Y., Saeed, N., Saha, A., Salem, M., Sanchez-Matilla, R., Schilling, K., Shao, W., Shen, Z., Shi, R., Shi, P., Sobotka, D., Soulier, T., Fadida, B.S., Stoyanov, D., Mun, T.S.H., Sun, X., Tao, R., Thaler, F., Théberge, A., Thielke, F., Torres, H., Wahid, K.A., Wang, J., Wang, Y., Wang, W., Wang, X., Wen, J., Wen, N., Wodzinski, M., Wu, Y., Xia, F., Xiang, T., Xiaofei, C., Xu, L., Xue, T., Yang, Y., Yang, L., Yao, K., Yao, H., Yazdani, A., Yip, M., Yoo, H., Yousefirizi, F., Yu, S., Yu, L., Zamora, J., Zeineldin, R.A., Zeng, D., Zhang, J., Zhang, B., Zhang, J., Zhang, F., Zhang, H., Zhao, Z., Zhao, Z., Zhao, J., Zhao, C., Zheng, Q., Zhi, Y., Zhou, Z., Zou, B., Maier-Hein, K., Jäger, P.F., Kopp-Schneider, A., Maier-Hein, L., 2022. Biomedical image analysis competitions: The state of current participation practice. URL: https://arxiv.org/abs/2212.08568, doi:10.48550/ARXIV.2212.08568.

Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K., 2021a. A large-scale study on unsupervised spatiotemporal representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3299–3309.

Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R.B., He, K., 2021b. A large-scale study on unsupervised spatiotemporal representation learning,

in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE. pp. 3299–3309. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Feichtenhofer_A_Large-Scale_Study_on_Unsupervised_Spatiotemporal_Representation_Learning_CVPR_2021_paper.html.

Fernando, B., Bilen, H., Gavves, E., Gould, S., 2017. Self-supervised video representation learning with odd-one-out networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society. pp. 5729–5738. URL: https://doi.org/10.1109/CVPR.2017.607, doi:10.1109/CVPR.2017.607.

Funke, I., Jenke, A., Mees, S.T., Weitz, J., Speidel, S., Bodenstedt, S., 2018. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis, in: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis. Springer, pp. 85–93.

Garrow, C.R., Kowalewski, K.F., Li, L., Wagner, M., Schmidt, M.W., Engelhardt, S., Hashimoto, D.A., Kenngott, H.G., Bodenstedt, S., Speidel, S., Müller-Stich, B.P., Nickel, F., 2021. Machine learning for surgical phase recognition: A systematic review. Annals of Surgery 273. doi:10.1097/SLA.0000000000004425.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 .

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al., 2021. Self-supervised pretraining of visual features in the wild. arXiv preprint arXiv:2103.01988 .

Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D., 2021. Cadis: Cataract dataset for surgical rgb-image segmentation. Medical Image Anal. 71, 102053. URL: https://doi.org/10.1016/j.media.2021.102053, doi:10.1016/j.media.2021.102053.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020a. Bootstrap your own latent - A new approach to self-supervised learning, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020b. Bootstrap your own latent: A new approach to self-supervised learning. CoRR abs/2006.07733. URL: https://arxiv.org/abs/2006.07733, arXiv:2006.07733.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. pp. 1735–1742.

Han, T., Xie, W., Zisserman, A., 2020. Self-supervised co-training for video representation learning, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/3def184ad8f4755ff269862ea77393dd-Abstract.html.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. arXiv:1911.05722.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR.

Henaff, O., 2020. Data-efficient image recognition with contrastive predictive coding, in: International Conference on Machine Learning, PMLR. pp. 4182–4192.

Hinton, G., Vinyals, O., Dean, J., et al., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2.

Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2018. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 .

Hu, X., Zeng, D., Xu, X., Shi, Y., 2021. Semi-supervised contrastive learning for label-efficient medical image segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 481–490. URL: https://doi.org/10.1007/978-3-030-87196-3_45, doi:10.1007/978-3-030-87196-3_45.

Huang, Y., Lin, L., Cheng, P., Lyu, J., Tang, X., 2021. Lesion-based contrastive learning for diabetic retinopathy grading from fundus images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 113–123. URL: https://doi.org/10.1007/978-3-030-87196-3_11, doi:10.1007/978-3-030-87196-3_11.

Jenni, S., Meishvili, G., Favaro, P., 2020. Video representation learning by recognizing temporal transformations, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII, Springer. pp. 425–442. URL: https://doi.org/10.1007/978-3-030-58604-1_26, doi:10.1007/978-3-030-58604-1_26.

Jiao, J., Cai, Y., Alsharid, M., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2020. Self-supervised contrastive video-speech representation learning for ultrasound, in: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part III, Springer. pp. 534–543. URL: https://doi.org/10.1007/978-3-030-59716-0_51, doi:10.1007/978-3-030-59716-0_51.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C., Heng, P., 2018. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Medical Imaging 37, 1114–1126. URL: https://doi.org/10.1109/TMI.2017.2787657.

Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C., Heng, P., 2020. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical Image Anal. 59. URL: https://doi.org/10.1016/j.media.2019.101572.

Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A., 2021. Temporal memory relation network for workflow recognition from surgical video. IEEE Transactions on Medical Imaging 40, 1911–1923. doi:10.1109/TMI.2021.3069471.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence .

Jing, L., Tian, Y., 2021. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43, 4037–4058. URL: https://doi.org/10.1109/TPAMI.2020.2992393, doi:10.1109/TPAMI.2020.2992393.

Kassem, H., Alapatt, D., Mascagni, P., Karargyris, A., Padoy, N., 2022. Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. IEEE Transactions on Medical Imaging .

Ke, J., Shen, Y., Liang, X., Shen, D., 2021. Contrastive learning based stain normalization across multiple tumor in histopathology, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VIII, Springer. pp. 571–580. URL: https://doi.org/10.1007/978-3-030-87237-3_55, doi:10.1007/978-3-030-87237-3_55.

Kim, D., Cho, D., Kweon, I.S., 2019. Self-supervised video representation learning with space-time cubic puzzles, in: AAAI 2019, AAAI Press. pp. 8545–8552. URL: https://doi.org/10.1609/aaai.v33i01.33018545, doi:10.1609/aaai.v33i01.33018545.

Kim, D., Cho, D., Yoo, D., Kweon, I.S., 2018. Learning image representations by completing damaged jigsaw puzzles, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 793–802.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization.

arXiv preprint arXiv:1412.6980 .

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T., 2011. HMDB: A large video database for human motion recognition, in: Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.V. (Eds.), IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, IEEE Computer Society. pp. 2556–2563. URL: https://doi.org/10.1109/ICCV.2011.6126543, doi:10.1109/ICCV.2011.6126543.

Lee, H., Huang, J., Singh, M., Yang, M., 2017. Unsupervised representation learning by sorting sequences, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society. pp. 667–676. URL: https://doi.org/10.1109/ICCV.2017.79, doi:10.1109/ICCV.2017.79.

Lei, W., Xu, W., Gu, R., Fu, H., Zhang, S., Zhang, S., Wang, G., 2021. Contrastive learning of relative position regression for one-shot object localization in 3d medical images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 155–165. URL: https://doi.org/10.1007/978-3-030-87196-3_15, doi:10.1007/978-3-030-87196-3_15.

Li, X., Ge, Y., Yi, K., Hu, Z., Shan, Y., Duan, L., 2022. mc-beit: Multi-choice discretization for image BERT pre-training, in: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX, Springer. pp. 231–246. URL: https://doi.org/10.1007/978-3-031-20056-4_14, doi:10.1007/978-3-031-20056-4_14.

Li, Z., Cui, Z., Wang, S., Qi, Y., Ouyang, X., Chen, Q., Yang, Y., Xue, Z., Shen, D., Cheng, J., 2021. Domain generalization for mammography detection via multi-style and multi-view contrastive learning, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VII, Springer. pp. 98–108. URL: https://doi.org/10.1007/978-3-030-87234-2_10, doi:10.1007/978-3-030-87234-2_10.

Liu, B., Zhan, L., Wu, X., 2021. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 210–220. URL: https://doi.org/10.1007/978-3-030-87196-3_20, doi:10.1007/978-3-030-87196-3_20.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., Nakawala, H., Park, A., Pugh, C.M., Stoyanov, D., Vedula, S.S., Cleary, K., Fichtinger, G., Forestier, G., Gibaud, B., Grantcharov, T.P., Hashizume, M., Heckmann-Nötzel, D., Kenngott, H.G., Kikinis, R., Mündermann, L., Navab, N., Onogur, S., Roß, T., Sznitman, R., Taylor, R.H., Tizabi, M.D., Wagner, M., Hager, G.D., Neumuth, T., Padoy, N., Collins, J., Gockel, I., Goedeke, J., Hashimoto, D.A., Joyeux, L., Lam, K., Leff, D.R., Madani, A., Marcus, H.J., Meireles, O.R., Seitel, A., Teber, D., Ückert, F., Müller-Stich, B.P., Jannin, P., Speidel, S., 2022. Surgical data science - from concepts toward clinical translation. Medical Image Anal. 76, 102306. URL: https://doi.org/10.1016/j.media.2021.102306, doi:10.1016/j.media.2021.102306.

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katić, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Jannin, P., 2017. Surgical data science for next-generation interventions. Nature Biomedical Engineering 1. doi:10.1038/s41551-017-0132-7.

Mascagni, P., Alapatt, D., Sestini, L., Altieri, M.S., Madani, A., Watanabe, Y., Alseidi, A., Redan, J.A., Alfieri, S., Costamagna, G., et al., 2022. Computer vision in surgery: from potential to clinical value. npj Digital Medicine 5, 1–9.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR. pp. 1273–1282.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al., 2017. Mixed precision training. arXiv preprint arXiv:1710.03740 .

Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717.

Misra, I., Zitnick, C.L., Hebert, M., 2016. Shuffle and learn: Unsupervised learning using temporal order verification, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, Springer. pp. 527–544. URL: https://doi.org/10.1007/978-3-319-46448-0_32, doi:10.1007/978-3-319-46448-0_32.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: European conference on computer vision, Springer. pp. 69–84.

Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al., 2022a. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. arXiv preprint arXiv:2204.04746 .

Nwoye, C.I., Mutter, D., Marescaux, J., Padoy, N., 2019. Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. International journal of computer assisted radiology and surgery 14, 1059–1067.

Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2022b. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Anal. 78, 102433. URL: https://doi.org/10.1016/j.media.2022.102433, doi:10.1016/j.media.2022.102433.

Van den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv e-prints , arXiv–1807.

van den Oord, A., Vinyals, O., Kavukcuoglu, K., 2017. Neural discrete representation learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 6309–6318.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Padoy, N., Blum, T., Ahmadi, S., Feußner, H., Berger, M., Navab, N., 2012. Statistical modeling and recognition of surgical workflow. Medical Image Anal. 16, 632–641. URL: https://doi.org/10.1016/j.media.2010.10.001, doi:10.1016/j.media.2010.10.001.

Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W., 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE. pp. 11205–11214. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Pan_VideoMoCo_Contrastive_Video_Representation_Learning_With_Temporally_Adversarial_Examples_CVPR_2021_paper.html.

Pathak, D., Girshick, R.B., Dollár, P., Darrell, T., Hariharan, B., 2017. Learning features by watching objects move, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society. pp. 6024–6033. URL: https://doi.org/10.1109/CVPR.2017.638, doi:10.1109/CVPR.2017.638.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society. pp. 2536–2544. URL: https://doi.org/10.1109/CVPR.2016.278, doi:10.1109/CVPR.2016.278.

Peng, C., Xiao, T., Li, Z., Jiang, Y., Zhang, X., Jia, K., Yu, G., Sun, J., 2018. Megdet: A large mini-batch object detector, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6181–6189.

Qian, R., Meng, T., Gong, B., Yang, M., Wang, H., Belongie, S.J., Cui, Y., 2021. Spatiotemporal contrastive video representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE. pp. 6964–6974. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Qian_Spatiotemporal_Contrastive_Video_Representation_Learning_CVPR_2021_paper.html.

Rivoir, D., Funke, I., Speidel, S., 2022. On the pitfalls of batch normalization for end-to-end video learning: A study on surgical workflow analysis.

This article has been accepted for publication in Medical Image Analysis. DOI: 10.1016/j.media.2023.102844 (2023)

URL: https://arxiv.org/abs/2203.07976, doi:10.48550/ARXIV.2203.07976.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al., 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery 13, 925–933.

Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N., 2021. A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. IEEE Robotics and Automation Letters 6, 2938–2945.

Shi, X., Jin, Y., Dou, Q., Heng, P., 2021. Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition. Medical Image Anal. 73, 102158. URL: https://doi.org/10.1016/j.media.2021.102158, doi:10.1016/j.media.2021.102158.

Soomro, K., Zamir, A.R., Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR abs/1212.0402. URL: http://arxiv.org/abs/1212.0402, arXiv:1212.0402.

Tian, Y., Krishnan, D., Isola, P., 2020. Contrastive multiview coding, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer. pp. 776–794.

Tian, Y., Pang, G., Liu, F., Chen, Y., Shin, S., Verjans, J.W., Singh, R., Carneiro, G., 2021. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part V, Springer. pp. 128–140. URL: https://doi.org/10.1007/978-3-030-87240-3_13, doi:10.1007/978-3-030-87240-3_13.

Twinanda, A.P., Mutter, D., Marescaux, J., Mathelin, M., Padoy, N., 2016a. Single- and multi-task architecture for surgical workflow at m2cai 2016. arXiv: Computer Vision and Pattern Recognition .

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016b. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging 36, 86–97.

Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K., 2018. Tracking emerges by colorizing videos, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII, Springer. pp. 402–419. URL: https://doi.org/10.1007/978-3-030-01261-8_24, doi:10.1007/978-3-030-01261-8_24.

Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., Reinke, A., Yu, T., Vardazaryan, A., Nwoye, C.I., Padoy, N., Liu, X., Lee, E., Disch, C., Meine, H., Xia, T., Jia, F., Kondo, S., Reiter, W., Jin, Y., Long, Y., Jiang, M., Dou, Q., Heng, P., Twick, I., Kirtaç, K., Hosgor, E., Bolmgren, J.L., Stenzel, M., von Siemens, B., Kenngott, H.G., Nickel, F., von Frankenberg, M., Mathis-Ullrich, F., Maier-Hein, L., Speidel, S., Bodenstedt, S., 2021. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. CoRR abs/2109.14956. URL: https://arxiv.org/abs/2109.14956, arXiv:2109.14956.

Wang, X., Jabri, A., Efros, A.A., 2019. Learning correspondence from the cycle-consistency of time, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE. pp. 2566–2576. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Learning_Correspondence_From_the_Cycle-Consistency_of_Time_CVPR_2019_paper.html, doi:10.1109/CVPR.2019.00267.

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. Dense contrastive learning for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3024–3033.

Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C., 2022. Masked feature prediction for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14668–14678.

Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J., 2021. Federated contrastive learning for volumetric medical image segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III, Springer. pp. 367–377. URL: https://doi.org/10.1007/978-3-030-87199-4_35, doi:10.1007/978-3-030-87199-4_35.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742.

Xiao, T., Wang, X., Efros, A.A., Darrell, T., 2020. What should not be contrastive in contrastive learning, in: International Conference on Learning Representations.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. Simmim: A simple framework for masked image modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9653–9663.

Xing, X., Hou, Y., Li, H., Yuan, Y., Li, H., Meng, M.Q., 2021. Categorical relation-preserving contrastive knowledge distillation for medical image classification, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part V, Springer. pp. 163–173. URL: https://doi.org/10.1007/978-3-030-87240-3_16, doi:10.1007/978-3-030-87240-3_16.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y., 2019. Self-supervised spatiotemporal learning via video clip order prediction, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE. pp. 10334–10343. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Xu_Self-Supervised_Spatiotemporal_Learning_via_Video_Clip_Order_Prediction_CVPR_2019_paper.html, doi:10.1109/CVPR.2019.01058.

Yang, H., Kahrs, L.A., 2021. Real-time coarse-to-fine depth estimation on stereo endoscopic images with self-supervised learning, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 733–737.

Yang, Y., Fang, H., Du, Q., Li, F., Zhang, X., Tan, M., Xu, Y., 2021. Distinguishing differences matters: Focal contrastive network for peripheral anterior synechiae recognition, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VIII, Springer. pp. 24–33. URL: https://doi.org/10.1007/978-3-030-87237-3_3, doi:10.1007/978-3-030-87237-3_3.

Yengera, G., Mutter, D., Marescaux, J., Padoy, N., 2018. Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of cnn-lstm networks. arXiv preprint arXiv:1805.08569 .

You, Y., Gitman, I., Ginsburg, B., 2017. Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 .

Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y., 2021. Positional contrastive learning for volumetric medical image segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 221–230. URL: https://doi.org/10.1007/978-3-030-87196-3_21, doi:10.1007/978-3-030-87196-3_21.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization, in: European conference on computer vision, Springer. pp. 649–666.

Zhang, R., Isola, P., Efros, A.A., 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1058–1067.

Zhao, Z., Yang, G., 2021. Unsupervised contrastive learning of radiomics and deep features for label-efficient tumor classification, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, Springer. pp. 252–261. URL: https://doi.org/10.1007/978-3-030-87196-3_24, doi:10.1007/978-3-030-87196-3_24.

Zhou, B., Liu, C., Duncan, J.S., 2021. Anatomy-constrained contrastive learning for synthetic segmentation without ground-truth, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part I, Springer. pp. 47–56. URL: https://doi.org/10.1007/978-3-030-87193-2_5, doi:10.1007/978-3-030-87193-2_5.

Zhuang, C., Zhai, A.L., Yamins, D., 2019. Local aggregation for unsupervised learning of visual embeddings, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6002–6012.

Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D., 2018. DeepPhase: Surgical phase recognition in cataracts videos, in: MICCAI.

Bibliography

[Acar 2025] A. Acar, J. Atoum, P. S. Connor, C. Pierre, C. N. Lynch, N. L. Kavoussi and J. Y. Wu. *NAVIUS: Navigated Augmented Reality Visualization for Ureteroscopic Surgery*. arXiv preprint, vol. arXiv:2503.17511, 2025. Accessed April 23 2025.

[Achanta 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. V. Fua and S. Süsstrunk. *SLIC Superpixels Compared to State-of-the-Art Superpixel Methods*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pages 2274–2282, 2012.

[Achiam 2023] J. Achiam. *GPT-4 Technical Report*. 2023.

[Agrawal 2015] P. Agrawal, J. Carreira and J. Malik. *Learning to See by Moving*. 2015 IEEE International Conference on Computer Vision (ICCV), pages 37–45, 2015.

[Allan 2019] M. Allan, A. A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, L. C. García-Peraza, W. Li, V. I. Iglovikov, H. Luo, J. Yang, D. Stoyanov, L. Maier-Hein, S. Speidel and M. Azizian. *2017 Robotic Instrument Segmentation Challenge*. ArXiv, vol. abs/1902.06426, 2019.

[Allan 2020] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes-Hurtado, E. Flouty, A. K. Mohammed, M. Pedersen, A. Kori, A. Varghese, G. Krishnamurthi, D. Rauber, R. Mendel, C. Palm, S. Bano, G. Saibro, C.-S. Shih, H.-A. Chiang, J. Zhuang, J. Yang, V. I. Iglovikov, A. Dobrenkii, M. Reddiboina, A. Reddy, X. Liu, C. Gao, M. Unberath, M. Azizian, D. Stoyanov, L. Maier-Hein and S. Speidel. *2018 Robotic Scene Segmentation Challenge*. ArXiv, vol. abs/2001.11190, 2020.

[Allen 2013] K. Allen. *How AI Is Changing The Way We Communicate: The Future Of Interaction*. Forbes, 2013.

[Anne-Sophie 2009] N. Anne-Sophie and B. Adelaide. *Verbal Communication as a sign of adaptation in socio-technical systems: the case of robotic surgery*. 2009.

[Arica 2024] S. Arica, O. Rubin, S. Gershov and S. Laufer. *Cuvler: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23105–23114, 2024.

[Arnab 2021] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic and C. Schmid. *ViViT: A Video Vision Transformer*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6816–6826, 2021.

[Asano 2020] Y. M. Asano, C. Rupprecht and A. Vedaldi. *A critical analysis of self-supervision, or what we can learn from a single image*. arXiv: CVPR, 2020.

[Avgousti 2020] S. Avgousti, E. G. Christoforou, A. S. Panayides, P. Masouras, P. Vieyres and C. S. Pattichis. *Robotic Systems in Current Clinical Practice*. 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), pages 269–274, 2020.

[Aytar 2016] Y. Aytar, C. Vondrick and A. Torralba. *SoundNet: Learning Sound Representations from Unlabeled Video*. NeurIPS, vol. abs/1610.09001, 2016.

[Baccouche 2011] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia and A. Baskurt. *Sequential Deep Learning for Human Action Recognition*. In International Workshop on Human Behavior Unterstanding, 2011.

[Baccouche 2012] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia and A. Baskurt. *Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification*. In British Machine Vision Conference, 2012.

[Bachmann 2022] R. Bachmann, D. Mizrahi, A. Atanov and A. Zamir. *MultiMAE: Multi-modal Multi-task Masked Autoencoders*. ECCV, vol. abs/2204.01678, 2022.

[Bao 2021] H. Bao, L. Dong, S. Piao and F. Wei. *Beit: Bert pre-training of image transformers*. arXiv preprint arXiv:2106.08254, 2021.

[Basiev 2021] K. Basiev, A. Goldbraikh, C. M. Pugh and S. Laufer. *Open surgery tool classification and hand utilization using a multi-camera system*. International Journal of Computer Assisted Radiology and Surgery, vol. 17, pages 1497 – 1505, 2021.

[Bastian 2023a] L. Bastian, T. Czempiel, C. Heiliger, K. Karcz, U. Eck, B. Busam and N. Navab. *Know your sensors—a modality study for surgical action classification*. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 11, no. 4, pages 1113–1121, 2023.

[Bastian 2023b] L. Bastian, D. Derkacz-Bogner, T. D. Wang, B. Busam and N. Navab. *SegmentOR: Obtaining Efficient Operating Room Semantics*

Through Temporal Propagation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 57–67. Springer, 2023.

[Bastian 2023c] L. Bastian, T. D. Wang, T. Czempiel, B. Busam and N. Navab. *DisguisOR: holistic face anonymization for the operating room*. International Journal of Computer Assisted Radiology and Surgery, vol. 18, pages 1209 – 1215, 2023.

[Batić 2024] D. Batić, F. Holm, E. Özsoy, T. Czempiel and N. Navab. *EndoViT: pretraining vision transformers on a large collection of endoscopic images*. International Journal of Computer Assisted Radiology and Surgery, vol. 19, pages 1085 – 1091, 2024.

[Baxter 2011] G. D. Baxter and I. Sommerville. *Socio-technical systems: From design methods to systems engineering*. Interact. Comput., vol. 23, pages 4–17, 2011.

[Bay 2008] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool. *Speeded-Up Robust Features (SURF)*. Comput. Vis. Image Underst., vol. 110, pages 346–359, 2008.

[Belagiannis 2016] V. Belagiannis, X. Wang, H. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. V. Fua, S. Ilic, H. Feußner and N. Navab. *Parsing human skeletons in an operating room*. Machine Vision and Applications, vol. 27, pages 1035–1046, 2016.

[Bertasius 2021] G. Bertasius, H. Wang and L. Torresani. *Is Space-Time Attention All You Need for Video Understanding?* ArXiv, vol. abs/2102.05095, 2021.

[Bewley 2016] A. Bewley, Z. Ge, L. Ott, F. T. Ramos and B. Upcroft. *Simple online and realtime tracking*. ICIP, 2016.

[Blum 2008] T. Blum, N. Padoy, H. Feußner and N. Navab. *Modeling and Online Recognition of Surgical Phases Using Hidden Markov Models*. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 11 Pt 2, pages 627–35, 2008.

[Bodenstedt 2017] S. Bodenstedt, M. Wagner, D. Katic, P. Mietkowski, B. F. B. Mayer, H. Kenngott, B. P. Müller-Stich, R. Dillmann and S. Speidel. *Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis*. ArXiv, vol. abs/1702.03684, 2017.

[Bollen 2011] J. Bollen, H. Mao and X. Zeng. *Twitter mood predicts the stock market*. Journal of computational science, vol. 2, no. 1, pages 1–8, 2011.

[Bosch 2002] F. Bosch, U. Wehrman, H. D. Saeger and W. Kirch. *Laparoscopic or open conventional cholecystectomy: clinical and economic considerations.* The European journal of surgery = Acta chirurgica, vol. 168 5, pages 270–7, 2002.

[Cai 2018] Z. Cai and N. Vasconcelos. *Cascade r-cnn: Delving into high quality object detection*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6154–6162, 2018.

[Cai 2023] D. Cai, Y. Kang, A. Yao and Y. Chen. *Ske2Grid: Skeleton-to-Grid Representation Learning for Action Recognition*. In International Conference on Machine Learning, 2023.

[Cai 2024] K. Cai. *Google brings AI answers to map applications*. Reuters, 2024.

[Cao 2017] Z. Cao, T. Simon, S.-E. Wei and Y. Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields*. In CVPR, pages 7291–7299, 2017.

[Carinou 2011] E. Carinou, M. Brodecki, J. Domienik, L. Donadille, C. Koukorava, S. Krim, D. Nikodemová, N. ruiz Lopez, M. Sans-Merce, L. Struelens and F. Vanhavere. *Recommendations to reduce extremity and eye lens doses in interventional radiology and cardiology*. Radiation Measurements, vol. 46, pages 1324–1329, 2011.

[Carion 2020a] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko. *End-to-End Object Detection with Transformers*. ArXiv, vol. abs/2005.12872, 2020.

[Carion 2020b] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko. *End-to-end object detection with transformers*. In European conference on computer vision, pages 213–229. Springer, 2020.

[Caron 2020] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski and A. Joulin. *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments*. NeurIPS, vol. abs/2006.09882, 2020.

[Caron 2021] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski and A. Joulin. *Emerging Properties in Self-Supervised Vision Transformers*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9630–9640, 2021.

[Carreira 2017] J. Carreira and A. Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017.

[Catchpole 2007] K. Catchpole, M. R. de Leval, A. I. McEwan, N. Pigott, M. Elliott, A. Mcquillan, C. MacDonald and A. J. Goldman. *Patient handover from surgery to intensive care: using Formula 1 pit-stop and aviation models to improve safety and quality*. Pediatric Anesthesia, vol. 17, 2007.

[Catchpole 2015] K. Catchpole, C. E. Perkins, C. Bresee, M. J. Solnik, B. Sherman, J. L. Fritch, B. Gross, S. Jagannathan, N. Hakami-Majd, R. M. Avenido and J. T. Anger. *Safety, efficiency and learning curves in robotic surgery: a human factors analysis*. Surgical Endoscopy, vol. 30, pages 3749–3761, 2015.

[Catchpole 2018] K. R. Catchpole, E. Hallett, S. Curtis, T. Mirchi, C. P. Souders and J. T. Anger. *Diagnosing barriers to safety and efficiency in robotic surgery*. Ergonomics, vol. 61, no. 1, pages 26–39, 2018.

[Catchpole 2024] K. Catchpole, T. Cohen, M. Alfred, S. Lawton, F. Kanji, D. Shouhed, L. Nemeth and J. Anger. *Human factors integration in robotic surgery*. Human factors, vol. 66, no. 3, pages 683–700, 2024.

[cCaughan Koksal 2024] cCaughan Koksal, G. Ghazaei, F. Holm, A. Farshad and N. Navab. *SANGRIA: Surgical Video Scene Graph Optimization for Surgical Workflow Prediction*. ArXiv, vol. abs/2407.20214, 2024.

[Chakraborty 2013] I. Chakraborty, A. Elgammal and R. S. Burd. *Video based activity recognition in trauma resuscitation*. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–8, 2013.

[Chaudhry 2009] R. A. Chaudhry, A. Ravichandran, G. Hager and R. Vidal. *Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1932–1939, 2009.

[Chen 2017] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam. *Rethinking atrous convolution for semantic image segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[Chen 2020] T. Chen, S. Kornblith, M. Norouzi and G. E. Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. ICML, 2020.

[Chen 2021a] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu. *Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 13339–13348, 2021.

[Chen 2021b] Z. Chen, S. Li, B. Yang, Q. Li and H. Liu. *Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition*. In AAAI Conference on Artificial Intelligence, 2021.

[Chen 2025] K. Chen, L. Schewski, V. K. Srivastav, J. L. Lavanchy, D. Mutter, G. Beldi, S. Keller and N. Padoy. *When do they StOP?: A First Step Towards Automatically Identifying Team Communication in the Operating Room*. ArXiv, vol. abs/2502.08299, 2025.

[Cheng] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng and H. Lu. *Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition*. In European Conference on Computer Vision.

[Cheng 2020] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng and H. Lu. *Skeleton-Based Action Recognition With Shift Graph Convolutional Network*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 180–189, 2020.

[Cherns 1976] A. B. Cherns. *The Principles of Sociotechnical Design*. Human Relations, vol. 29, pages 783 – 792, 1976.

[Childers 2018] C. P. Childers and M. Maggard-Gibbons. *Understanding costs of care in the operating room*. JAMA surgery, vol. 153, no. 4, pages e176233–e176233, 2018.

[Chung 2014] J. Chung, Çaglar Gülçehre, K. Cho and Y. Bengio. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. ArXiv, vol. abs/1412.3555, 2014.

[Cofran 2021] L. Cofran, T. Cohen, M. Alfred, F. Kanji, E. Choi, S. Savage, J. Anger and K. Catchpole. *Barriers to safety and efficiency in robotic surgery docking*. Surgical endoscopy, pages 1–10, 2021.

[Czempiel 2020] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feußner, S. T. Kim and N. Navab. *TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks*. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020.

[Czempiel 2021] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam and N. Navab. *OperA: Attention-Regularized Transformers for Surgical Phase Recognition*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021.

[Dai 2024] P.-Y. Dai, Y.-C. Wu, R.-K. Sheu, C.-L. Wu, S.-F. Liu, P.-Y. Lin, W.-L. Cheng, G.-Y. Lin, H.-C. Chung and L.-C. Chen. *An automated ICU agitation monitoring system for video streaming using deep learning classification*. BMC Medical Informatics and Decision Making, vol. 24, 2024.

[Davoudi 2019] A. Davoudi, K. R. Malhotra, B. Shickel, S. Siegel, S. Williams, M. M. Ruppert, E. Bihorac, T. Ozrazgat-Baslanti, P. J. Tighe, A. Bihorac and P. Rashidi. *Intelligent ICU for Autonomous Patient Monitoring Using Pervasive Sensing and Deep Learning*. Scientific Reports, vol. 9, 2019.

[Devlin 2019] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In North American Chapter of the Association for Computational Linguistics, 2019.

[Dias 2020] R. D. Dias, S. J. Yule and M. A. Zenati. *Augmented Cognition in the Operating Room*. 2020.

[Ding 2021] D. Ding, F. Hill, A. Santoro and M. M. Botvinick. *Attention over learned object embeddings enables complex visual reasoning*. NeurIPS, 2021.

[Do 2024] J. Do and M. Kim. *SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition*. ArXiv, vol. abs/2403.09508, 2024.

[Doersch 2015] C. Doersch, A. K. Gupta and A. A. Efros. *Unsupervised Visual Representation Learning by Context Prediction*. 2015 IEEE ICCV, 2015.

[Donahue 2014] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell. *Long-term recurrent convolutional networks for visual recognition and description*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2625–2634, 2014.

[Dosovitskiy 2020] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ArXiv, vol. abs/2010.11929, 2020.

[Du 2015] Y. Du, W. Wang and L. Wang. *Hierarchical recurrent neural network for skeleton based action recognition*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1110–1118, 2015.

[Duan 2023] H. Duan, M. Xu, B. Shuai, D. Modolo, Z. Tu, J. Tighe and A. Bergamo. *SkeleTR: Towards Skeleton-based Action Recognition in the Wild*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 13588–13598, 2023.

[Fallon 2024] A. Fallon, K. Haralambides, J. Mazzillo and C. Gleber. *Addressing Alert Fatigue by Replacing a Burdensome Interruptive Alert with Passive Clinical Decision Support*. Applied Clinical Informatics, vol. 15, no. 1, pages 101–110, 2024.

[Fan 2021] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik and C. Feichtenhofer. *Multiscale Vision Transformers*. ICCV, vol. abs/2104.11227, 2021.

[Farha 2019] Y. A. Farha and J. Gall. *MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3570–3579, 2019.

[Felzenszwalb 2004] P. F. Felzenszwalb and D. P. Huttenlocher. *Efficient graph-based image segmentation*. International journal of computer vision, vol. 59, pages 167–181, 2004.

[Flouty 2018] E. Flouty, O. Zisimopoulos and D. Stoyanov. *FaceOff: Anonymizing Videos in the Operating Rooms*. In OR 2.0/CARE/-CLIP/ISIC@MICCAI, 2018.

[Fujii 2024] R. Fujii, M. Hatano, H. Saito and H. Kajita. *EgoSurgery-Phase: A Dataset of Surgical Phase Recognition from Egocentric Open Surgery Videos*. ArXiv, vol. abs/2405.19644, 2024.

[Gao 2021] X. Gao, Y. Jin, Y. Long, Q. Dou and P.-A. Heng. *Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021.

[Gao 2022] Z. Gao, P. Wang, P. Lv, X. Jiang, Q. dong Liu, P. Wang, M. Xu and W. Li. *Focal and Global Spatial-Temporal Transformer for Skeleton-based Action Recognition*. ArXiv, vol. abs/2210.02693, 2022.

[Geng 2023] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li and H. Hu. *Human Pose as Compositional Tokens*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 660–671, 2023.

[Gershov 2022] S. Gershov, F. Mahameed, A. Raz and S. Laufer. *More Than Meets the Eye: Anesthesiologists' Visual Attention in the Operating Room*. In AMAI, 2022.

[Gidaris 2018] S. Gidaris, P. Singh and N. Komodakis. *Unsupervised Representation Learning by Predicting Image Rotations*. ICLR, 2018.

[Goel 2022] S. Goel, H. Bansal, S. K. Bhatia, R. A. Rossi, V. Vinay and A. Grover. *CyCLIP: Cyclic Contrastive Language-Image Pretraining*. ArXiv, vol. abs/2205.14459, 2022.

[Goyal 2017] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. N. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax and R. Memisevic. *The "Something Something" Video Database for Learning and Evaluating Visual Common Sense*. 2017 IEEE International Conference on Computer Vision (ICCV), pages 5843–5851, 2017.

[Grill 2020] J.-B. Grill, F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos and M. Valko. *Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning*. NeurIPS, vol. abs/2006.07733, 2020.

[Gupta 2015] S. Gupta, J. Hoffman and J. Malik. *Cross Modal Distillation for Supervision Transfer*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2827–2836, 2015.

[Hajj 2019] H. A. Hajj, M. Lamard, P.-H. Conze, S. Roychowdhury, X. Hu, G. Marsalkaite, O. Zisimopoulos, M. A. Dedmari, F. Zhao, J. Prellberg, M. Sahu, A. Galdran, T. Araújo, D. M. Vo, C. K. Panda, N. Dahiya, S. Kondo, Z. Bian and G. Quellec. *CATARACTS: Challenge on automatic tool annotation for cataRACT surgery*. Medical Image Analysis, vol. 52, page 24–41, 2019.

[Hamoud 2022] I. Hamoud, A. Karargyris, A. Sharghi, O. Mohareri and N. Padoy. *Self-supervised learning via cluster distance prediction for operating room context awareness*. IJCARS, 2022.

[Hamoud 2023] I. Hamoud, M. A. Jamal, V. Srivastav, D. Mutter, N. Padoy and O. Mohareri. *ST(OR)^2: Spatio-Temporal Object Level Reasoning for Activity Recognition in the Operating Room*. In International Conference on Medical Imaging with Deep Learning, MIDL 2023, 10-12 July

2023, Nashville, USA, Proceedings of Machine Learning Research. PMLR, 2023.

[Hamoud 2025] I. Hamoud, V. K. Srivastav, M. A. Jamal, D. Mutter, O. Mohareri and N. Padoy. *Multi-view Video-Pose Pretraining for Operating Room Surgical Activity Recognition*. ArXiv, vol. abs/2502.13883, 2025.

[Haro 2012] B. B. Haro, L. Zappella and R. Vidal. *Surgical Gesture Classification from Video Data*. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 15 Pt 1, pages 34–41, 2012.

[Hatano 2024] M. Hatano, R. Hachiuma, R. Fujii and H. Saito. *Multimodal Cross-Domain Few-Shot Learning for Egocentric Action Recognition*. ECCV, 2024.

[He 2015] K. He, X. Zhang, S. Ren and J. Sun. *Deep Residual Learning for Image Recognition*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2015.

[He 2017] K. He, G. Gkioxari, P. Dollár and R. B. Girshick. *Mask R-CNN*. In International Conference on Computer Vision, 2017.

[He 2019] K. He, H. Fan, Y. Wu, S. Xie and R. B. Girshick. *Momentum Contrast for Unsupervised Visual Representation Learning*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9726–9735, 2019.

[He 2022a] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick. *Masked autoencoders are scalable vision learners*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022.

[He 2022b] Z. He, A. Mottaghi, A. Sharghi, M. A. Jamal and O. Mohareri. *An Empirical Study on Activity Recognition in Long Surgical Videos*. In Proceedings of the 2nd Machine Learning for Health symposium, pages 356–372, 2022.

[Herzig 2022] R. Herzig, E. Ben-Avraham, K. Mangalam, A. Bar, G. Chechik, A. Rohrbach, T. Darrell and A. Globerson. *Object-Region Video Transformers*. In CVPR 2022, pages 3138–3149. IEEE, 2022.

[Hinton 2015] G. E. Hinton, O. Vinyals and J. Dean. *Distilling the Knowledge in a Neural Network*. ArXiv, 2015.

[Hong 2020] W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W. Chang and C.-S. Shih. *CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80*. ArXiv, vol. abs/2012.12453, 2020.

[Issenhuth 2019] T. Issenhuth, V. Srivastav, A. Gangi and N. Padoy. *Face detection in the operating room: Comparison of state-of-the-art methods and a self-supervised approach*. International journal of computer assisted radiology and surgery, vol. 14, pages 1049–1058, 2019.

[Jain 1996] A. K. Jain and A. Vailaya. *Image retrieval using color and shape*. Pattern Recognit., vol. 29, pages 1233–1244, 1996.

[Jamal 2022] M. A. Jamal and O. Mohareri. *Multi-Modal Unsupervised Pre-Training for Surgical Operating Room Workflow Analysis*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022.

[Jamal 2023a] M. A. Jamal and O. Mohareri. *M33D: Learning 3D priors using Multi-Modal Masked Autoencoders for 2D image and video understanding*. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2532–2542, 2023.

[Jamal 2023b] M. A. Jamal and O. Mohareri. *SurgMAE: Masked Autoencoders for Long Surgical Video Analysis*. ArXiv, vol. abs/2305.11451, 2023.

[Jia 2021] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li and T. Duerig. *Scaling up visual and vision-language representation learning with noisy text supervision*. In ICML, pages 4904–4916. PMLR, 2021.

[Johansson 1973] G. Johansson. *Visual perception of biological motion and a model for its analysis*. Perception & Psychophysics, vol. 14, pages 201–211, 1973.

[Johnston 2009] G. H. Johnston, L. Ekert and E. Pally. *Surgical site signing and "time out": issues of compliance or complacence.* The Journal of bone and joint surgery. American volume, vol. 91 11, pages 2577–80, 2009.

[Kadkhodamohammadi 2014] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Temporally Consistent 3D Pose Estimation in the Interventional Room Using Discrete MRF Optimization over RGBD Sequences*. In International Conference on Information Processing in Computer-Assisted Interventions, 2014.

[Kadkhodamohammadi 2015] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Pictorial Structures on RGB-D Images*

for Human Pose Estimation in the Operating Room. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015.

[Kahneman 1992] D. Kahneman, A. Treisman and B. J. Gibbs. *The reviewing of object files: Object-specific integration of information*. Cognitive Psychology, vol. 24, pages 175–219, 1992.

[Kanji 2021] F. Kanji, T. Cohen, M. Alfred, A. Caron, S. Lawton, S. Savage, D. Shouhed, J. T. Anger and K. Catchpole. *Room size influences flow in robotic-assisted surgery*. International Journal of Environmental Research and Public Health, vol. 18, no. 15, page 7984, 2021.

[Kanji 2024] F. F. Kanji, A. Marselian, M. Burch, M. Jain and T. N. Cohen. *Challenges With Robot-Assisted Surgery Setup for Complex Minimally Invasive Upper Gastrointestinal Surgery*. The American Surgeon™, vol. 90, no. 10, pages 2403–2410, 2024.

[Karpathy 2014] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei. *Large-Scale Video Classification with Convolutional Neural Networks*. 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[Ke 2010] Y. Ke, R. Sukthankar and M. Hebert. *Volumetric Features for Video Event Detection*. International Journal of Computer Vision, vol. 88, pages 339–362, 2010.

[Keller 2022] S. Keller, F. Tschan, N. K. Semmer, S. Trelle, T. Manser and G. Beldi. *StOP? II trial: cluster randomized clinical trial to test the implementation of a toolbox for structured communication in the operating room—study protocol*. Trials, vol. 23, 2022.

[Kennedy-Metz 2021] L. R. Kennedy-Metz, P. Mascagni, A. Torralba, R. D. Dias, P. Perona, J. A. Shah, N. Padoy and M. A. Zenati. *Computer Vision in the Operating Room: Opportunities and Caveats*. IEEE Transactions on Medical Robotics and Bionics, vol. 3, pages 2–10, 2021.

[Kipf 2016] T. Kipf and M. Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. ArXiv, vol. abs/1609.02907, 2016.

[Koch 2022] A. Koch, B. Schlenker, A. Becker and M. Weigl. *Operating room team strategies to reduce flow disruptions in high-risk task episodes: resilience in robot-assisted surgery*. Ergonomics, vol. 0, no. 0, pages 1–14, 2022.

[Korkiakangas 2014] T. Korkiakangas, S. M. Weldon, J. Bezemer and R. L. Kneebone. *Nurse-surgeon object transfer: video analysis of communication and situation awareness in the operating theatre.* International journal of nursing studies, vol. 51 9, pages 1195–206, 2014.

[Koschmann 2011] T. Koschmann, C. LeBaron, C. Goodwin and P. J. Feltovich. *Can you see the cystic artery yet? A simple matter of trust*. Journal of Pragmatics, vol. 43, pages 521–541, 2011.

[Koukorava 2014] C. Koukorava, J. Farah, L. Struelens, I. Clairand, L. Donadille, F. Vanhavere and P. Dimitriou. *Efficiency of radiation protection equipment in interventional radiology: a systematic Monte Carlo study of eye lens and whole body doses*. Journal of Radiological Protection, vol. 34, pages 509 – 528, 2014.

[Krebs 2022] A. Krebs, J.-P. Mazellier, J. M. Verde, C. Rolland, J. Bert and N. Padoy. *Organ-based estimation and minimization of clinician's X-ray dose*. International Journal of Computer Assisted Radiology and Surgery, vol. 17, pages 2357–2364, 2022.

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. Communications of the ACM, vol. 60, pages 84 – 90, 2012.

[Krähenbühl 2011] P. Krähenbühl and V. Koltun. *Efficient inference in fully connected CRFs with Gaussian edge potentials*. In Advances in neural information processing systems, volume 24, pages 109–117, 2011.

[Kuehne 2011] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio and T. Serre. *HMDB: A large video database for human motion recognition*. 2011 International Conference on Computer Vision, pages 2556–2563, 2011.

[Ladikos 2008] A. Ladikos, S. Benhimane and N. Navab. *Real-Time 3D Reconstruction for Collision Avoidance in Interventional Environments*. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 11 Pt 2, pages 526–34, 2008.

[Ladikos 2010] A. Ladikos, C. Cagniart, R. Ghotbi, M. Reiser and N. Navab. *Estimating Radiation Exposure in Interventional Environments*. MICCAI, vol. 13 Pt 3, 2010.

[Lalys 2010] F. Lalys, L. Riffaud, X. Morandi and P. Jannin. *Surgical Phases Detection from Microscope Videos by Combining SVM and HMM*. In MCV, 2010.

[Laptev 2003] I. Laptev and T. Lindeberg. *Space-time interest points*. Proceedings Ninth IEEE International Conference on Computer Vision, pages 432–439 vol.1, 2003.

[Lavanchy 2023a] J. L. Lavanchy, S. Ramesh, D. Dall'Alba, C. Gonzalez, P. Fiorini, B. P. Müller-Stich, P. C. Nett, J. Marescaux, D. Mutter and N. Padoy. *Challenges in multi-centric generalization: phase and step recognition in Roux-en-Y gastric bypass surgery*. International Journal of Computer Assisted Radiology and Surgery, vol. 19, pages 2249 – 2257, 2023.

[Lavanchy 2023b] J. L. Lavanchy, A. Vardazaryan, P. Mascagni, A. Consortium, D. Mutter and N. Padoy. *Preserving privacy in surgical video analysis using a deep learning classifier to identify out-of-body scenes in endoscopic videos*. Scientific Reports, vol. 13, 2023.

[Lee 2019] D. J. Lee, J. M. Ding and T. J. Guzzo. *Improving Operating Room Efficiency*. Current Urology Reports, vol. 20, pages 1–8, 2019.

[Lefkowitz 2018] M. Lefkowitz. *Study explores how robots in the operating room impact teamwork*. Cornell Chronicle, 2018.

[Li 2020a] Z. Li, A. Shaban, J.-G. Simard, D. Rabindran, S. DiMaio and O. Mohareri. *A robotic 3d perception system for operating room environment awareness*. arXiv preprint arXiv:2003.09487, 2020.

[Li 2020b] Z. Li, A. Shaban, J.-G. Simard, D. Rabindran, S. P. DiMaio and O. Mohareri. *A Robotic 3D Perception System for Operating Room Environment Awareness*. IPCAI, 2020.

[Li 2021] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu and J. Yan. *Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm*. ICLR, 2021.

[Li 2024] S. Li, J. Wang, R. Dai, W. Ma, W. Y. Ng, Y. Hu and Z. Li. *RoboNurse-VLA: Robotic Scrub Nurse System based on Vision–Language–Action Model*, 2024.

[Liang 2022] P. P. Liang, A. Zadeh and L.-P. Morency. *Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions*. ACM Computing Surveys, vol. 56, pages 1 – 42, 2022.

[Lin 2023] W. Lin, L. Karlinsky, N. Shvetsova, H. Possegger, M. Kozinski, R. Panda, R. Feris, H. Kuehne and H. Bischof. *Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge*. In ICCV, pages 2851–2862, 2023.

[Lindenberger 2023] P. Lindenberger, P.-E. Sarlin and M. Pollefeys. *LightGlue: Local Feature Matching at Light Speed*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 17581–17592, 2023.

[Liu 2022] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie. *A convnet for the 2020s*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11976–11986, 2022.

[Liu 2023] Y. Liu, M. Boels, L. C. García-Peraza-Herrera, T. K. M. Vercauteren, P. Dasgupta, A. Granados and S. Ourselin. *LoViT: Long Video Transformer for surgical phase recognition*. Medical Image Analysis, vol. 99, 2023.

[Loshchilov 2017] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. In ICLR, 2017.

[Lowe 2004] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer Vision, vol. 60, pages 91–110, 2004.

[Lu 2019] J. Lu, D. Batra, D. Parikh and S. Lee. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. In Neural Information Processing Systems, 2019.

[Macario 2006] A. Macario. *Are your hospital operating rooms" efficient"? A scoring system with eight performance indicators.* Anesthesiology, vol. 105, no. 2, pages 237–240, 2006.

[Maier-Hein 2022] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni*et al. Surgical data science–from concepts toward clinical translation*. Medical image analysis, vol. 76, page 102306, 2022.

[Manjunath 1996] B. S. Manjunath and W.-Y. Ma. *Texture Features for Browsing and Retrieval of Image Data*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, pages 837–842, 1996.

[Materzynska 2019] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang and T. Darrell. *Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks*. CVPR, vol. abs/1912.09930, 2019.

[McKinzie 2023] B. McKinzie, V. Shankar, J. Cheng, Y. Yang, J. Shlens and A. Toshev. *Robustness in Multimodal Learning under Train-Test Modality Mismatch*. In International Conference on Machine Learning, 2023.

[Misra 2016] I. Misra, C. L. Zitnick and M. Hebert. *Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification*. In European Conference on Computer Vision, 2016.

[Mizrahi 2024] D. Mizrahi, R. Bachmann, O. F. Kar, T. Yeo, M. Gao, A. Dehghan and A. Zamir. *4M: Massively Multimodal Masked Modeling*. NeurIPS, vol. abs/2312.06647, 2024.

[Moencks 2022] M. Moencks, E. Roth, T. Bohné, M. Basso and F. Betti. *Augmented Workforce: Empowering People, Transforming Manufacturing*. Technical report, World Economic Foru,, 2022.

[Mostafa 2025] M. L. Mostafa, A. Alperovich, D. Fedotov, G. Ghazaei, S. Saur, A. Farshad and N. Navab. *Surgical Flow Masked Autoencoder for Event Recognition*. In Medical Imaging with Deep Learning, 2025.

[Murali 2022] A. Murali, D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, D. Mutter and N. Padoy. *Latent Graph Representations for Critical View of Safety Assessment*. IEEE Transactions on Medical Imaging, vol. 43, pages 1247–1258, 2022.

[Murali 2023] A. Murali, D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, G. Costamagna, D. Mutter, J. Marescaux, B. Dallemagne and N. Padoy. *The Endoscapes Dataset for Surgical Scene Segmentation, Object Detection, and Critical View of Safety Assessment: Official Splits and Benchmark*. ArXiv, vol. abs/2312.12429, 2023.

[Nespolo 2022] R. Nespolo, E. D. Cole, D. Wang, D. Yi and Y. I. Leiderman. *A Platform for Tracking Surgeon and Observer Gaze as a Surrogate for Attention in Ophthalmic Surgery*. Ophthalmology Science, vol. 3, 2022.

[Newell 2020] A. Newell and J. Deng. *How Useful Is Self-Supervised Pretraining for Visual Tasks?* 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[OECD 2021] OECD. *Key Insights Proposed Solutions From the Future of Care and the Caregiving Workforce: Lessons and Insights from the COVID-19 Experience*. 2021.

[Özsoy 2022] E. Özsoy, E. P. Örnek, U. Eck, T. Czempiel, F. Tombari and N. Navab. *4D-OR: Semantic Scene Graphs for OR Domain Modeling*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022.

[Özsoy 2023] E. Özsoy, T. Czempiel, F. Holm, C. Pellegrini and N. Navab. *LABRAD-OR: Lightweight Memory Scene Graphs for Accurate Bimodal Reasoning in Dynamic Operating Rooms*. In International Conference

on Medical Image Computing and Computer-Assisted Intervention, 2023.

[Özsoy 2025] E. Özsoy, C. Pellegrini, T. Czempiel, F. Tristram, K. Yuan, D. Bani-Harouni, U. Eck, B. Busam, M. Keicher and N. Navab. *MM-OR: A Large Multimodal Operating Room Dataset for Semantic Understanding of High-Intensity Surgical Environments*. ArXiv, vol. abs/2503.02579, 2025.

[Padoy 2008] N. Padoy, T. Blum, H. Feußner, M. Berger and N. Navab. *On-line Recognition of Surgical Activity for Monitoring in the Operating Room*. In AAAI, 2008.

[Padoy 2009] N. Padoy, D. Mateus, D. Weinland, M.-O. Berger and N. Navab. *Workflow monitoring based on 3D motion features*. 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pages 585–592, 2009.

[Padoy 2012] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feußner, M.-O. Berger and N. Navab. *Statistical modeling and recognition of surgical workflow*. Medical image analysis, vol. 16 3, pages 632–41, 2012.

[Papandreou 2017] G. Papandreou, T. L. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler and K. P. Murphy. *Towards Accurate Multi-person Pose Estimation in the Wild*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3711–3719, 2017.

[Paszke 2019] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. ArXiv, vol. abs/1912.01703, 2019.

[Pathak 2016] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell and A. A. Efros. *Context Encoders: Feature Learning by Inpainting*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2536–2544, 2016.

[Patrick 2010] J. Patrick and P. L. Morgan. *Approaches to understanding, analysing and developing situation awareness*. Theoretical Issues in Ergonomics Science, vol. 11, no. 1-2, pages 41–57, 2010.

[Patrick 2021] M. Patrick. *Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers*. 2021.

[Press 2013] G. Press. *A Very Short History of Big Data*. Forbes, 2013.

[Qian 2024] R. Qian, S. Ding and D. Lin. *Rethinking Image-to-Video Adaptation: An Object-centric Perspective*. ArXiv, vol. abs/2407.06871, 2024.

[Rabiner 2007] L. R. Rabiner and R. W. Schafer. *Introduction to Digital Speech Processing*. Found. Trends Signal Process., vol. 1, pages 1–194, 2007.

[Radevski 2021] G. Radevski, M.-F. Moens and T. Tuytelaars. *Revisiting spatio-temporal layouts for compositional action recognition*. In British Machine Vision Conference, 2021.

[Radevski 2023] G. Radevski, D. Grujicic, M.-F. Moens, M. B. Blaschko and T. Tuytelaars. *Multimodal Distillation for Egocentric Action Recognition*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 5190–5201, 2023.

[Radford 2021] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. In International Conference on Machine Learning, 2021.

[Ramesh 2022] S. Ramesh, V. K. Srivastav, D. Alapatt, T. Yu, A. Murali, L. Sestini, C. I. Nwoye, I. Hamoud, A. Fleurentin, G. Exarchakis, A. Karargyris and N. Padoy. *Dissecting Self-Supervised Learning Methods for Surgical Computer Vision*. Medical image analysis, vol. 88, page 102844, 2022.

[Reddy 2023] K. Reddy, P. Gharde, H. A. Tayade, M. Patil, L. srivani Reddy and D. P. Surya. *Advancements in Robotic Surgery: A Comprehensive Overview of Current Utilizations and Upcoming Frontiers*. Cureus, vol. 15, 2023.

[Reid 2024] M. Reid. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. ArXiv, vol. abs/2403.05530, 2024.

[Ren 2016] S. Ren, K. He, R. Girshick and J. Sun. *Faster R-CNN: Towards real-time object detection with region proposal networks*. IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pages 1137–1149, 2016.

[Rivero-Moreno 2024] Y. Rivero-Moreno, M. Rodriguez, P. Losada-Muñoz, S. Redden, S. Lopez-Lezama, A. Vidal-Gallardo, D. Machado-Paled, J. Cordova Guilarte and S. Teran-Quintero. *Autonomous Robotic Surgery: Has the Future Arrived?* Cureus, vol. 16, no. 1, page e52243, 2024.

[Rodas 2017] N. L. Rodas, J. Bert, D. Visvikis, M. de Mathelin and N. Padoy. *Pose optimization of a C-arm imaging device to reduce intraoperative radiation exposure of staff and patient during interventional procedures*. 2017 IEEE International Conference on Robotics and Automation (ICRA), pages 4200–4207, 2017.

[Rodas 2018] N. L. Rodas and N. Padoy. *Augmented Reality for Reducing Intraoperative Radiation Exposure to Patients and Clinicians during X-Ray Guided Procedures*. Mixed and Augmented Reality in Medicine, 2018.

[Rosen 2018] M. A. Rosen, D. DiazGranados, A. S. Dietz, L. E. Benishek, D. Thompson, P. J. Pronovost and S. J. Weaver. *Teamwork in healthcare: Key discoveries enabling safer, high-quality care.* American Psychologist, vol. 73, no. 4, page 433, 2018.

[Ross 2017] T. Ross, D. Zimmerer, A. S. Vemuri, F. Isensee, S. Bodenstedt, F. Both, P. Kessler, M. Wagner, B. P. Müller-Stich, H. Kenngott, S. Speidel, K. H. Maier-Hein and L. Maier-Hein. *Exploiting the potential of unlabeled endoscopic video data with self-supervised learning*. International Journal of Computer Assisted Radiology and Surgery, vol. 13, pages 925–933, 2017.

[Roß 2018] T. Roß, D. Zimmerer, A. S. Vemuri, F. Isensee, S. Bodenstedt, F. Both, P. Kessler, M. Wagner, B. P. Müller-Stich, H. Kenngott, S. Speidel, K. Maier-Hein and L. Maier-Hein. *Exploiting the potential of unlabeled endoscopic video data with self-supervised learning*. IJCARS, 2018.

[Sakamoto 2024] T. Sakamoto, Y. Harada and T. Shimizu. *Facilitating Trust Calibration in Artificial Intelligence–Driven Diagnostic Decision Support Systems for Determining Physicians' Diagnostic Accuracy: Quasi-Experimental Study*. JMIR Formative Research, vol. 8, page e58666, 2024.

[Satava 2003] R. M. Satava. *The Operating Room of the Future: Observations and Commentary*. Surgical Innovation, vol. 10, pages 105 – 99, 2003.

[Satyanaik 2024] S. Satyanaik, A. Murali, D. Alapatt, X. Wang, P. Mascagni and N. Padoy. *Optimizing latent graph representations of surgical scenes for unseen domain generalization*. International journal of computer assisted radiology and surgery, 2024.

[Schiff 2016] L. Schiff, Z. Tsafrir, J. Aoun, A. R. Taylor, E. Theoharis and D. Eisenstein. *Quality of Communication in Robotic Surgery and Surgical Outcomes*. JSLS : Journal of the Society of Laparoendoscopic Surgeons, vol. 20, 2016.

[Schmidt 2021] A. Schmidt, A. Sharghi, H. Haugerud, D. Oh and O. Mohareri. *Multi-view Surgical Video Action Detection via Mixed Global View Attention*. In MICCAI, 2021.

[Scholl 2007] B. J. Scholl. *Object Persistence in Philosophy and Psychology*. Mind & Language, vol. 22, pages 563–591, 2007.

[Sharghi 2020] A. Sharghi, H. Haugerud, D. Oh and O. Mohareri. *Automatic Operating Room Surgical Activity Recognition for Robot-Assisted Surgery*. In MICCAI, 2020.

[Sheetz KH 2020] D. J. Sheetz KH Claflin J. *Trends in the Adoption of Robotic Surgery for Common Surgical Procedures*. JAMA Network Open, 2020.

[Shelhamer 2014] E. Shelhamer, J. Long and T. Darrell. *Fully convolutional networks for semantic segmentation*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3431–3440, 2014.

[Shi 2020] L. Shi, Y. Zhang, J. Cheng and H. Lu. *Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition*. In Asian Conference on Computer Vision, 2020.

[Siméoni 2021] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet and J. Ponce. *Localizing objects with self-supervised transformers and no labels*. arXiv preprint arXiv:2109.14279, 2021.

[Simonyan 2014] K. Simonyan and A. Zisserman. *Two-Stream Convolutional Networks for Action Recognition in Videos*. NeurIPS, vol. abs/1406.2199, 2014.

[Soomro 2012] K. Soomro, A. Zamir and M. Shah. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. ArXiv, vol. abs/1212.0402, 2012.

[Souders 2017] C. P. Souders, K. Catchpole, L. N. Wood, J. M. Solnik, R. Avenido, P. L. Strauss, K. S. Eilber and J. T. Anger. *Reducing Operating Room Turnover Time for Robotic Surgery Using a Motor Racing Pit Stop Model*. World Journal of Surgery, vol. 41, pages 1943–1949, 2017.

[Srivastav 2018] V. K. Srivastav, T. Issenhuth, A. Kadkhodamohammadi, M. de Mathelin, A. Gangi and N. Padoy. *MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation*. ArXiv, vol. abs/1808.08180, 2018.

[Srivastav 2020] V. K. Srivastav, A. Gangi and N. Padoy. *Self-supervision on Unlabelled OR Data for Multi-person 2D/3D Human Pose Estimation*. ArXiv, vol. abs/2007.08354, 2020.

[Srivastav 2021] V. K. Srivastav, A. Gangi and N. Padoy. *Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the OR*. Medical image analysis, vol. 80, page 102525, 2021.

[Sun 2022] X. Sun, P. Chen, L.-C. Chen, C. Li, T. H. Li, M. Tan and C. Gan. *Masked Motion Encoding for Self-Supervised Video Representation Learning*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2235–2245, 2022.

[Svensson 2007] M. S. Svensson, C. Heath and P. Luff. *Instrumental action: the timely exchange of implements during surgical operations*. In European Conference on Computer Supported Cooperative Work, 2007.

[Svensson 2009] M. S. Svensson, P. Luff and C. Heath. *Embedding instruction in practice: contingency and collaboration during surgical training.* Sociology of health & illness, vol. 31 6, pages 889–906, 2009.

[Taleb 2020] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner and C. Lippert. *3D Self-Supervised Methods for Medical Imaging*. NeurIPS, vol. abs/2006.03829, 2020.

[Taylor 2010] G. W. Taylor, R. Fergus, Y. LeCun and C. Bregler. *Convolutional Learning of Spatio-temporal Features*. In European Conference on Computer Vision, 2010.

[Tenenbaum 2011] J. B. Tenenbaum, C. Kemp, T. L. Griffiths and N. D. Goodman. *How to Grow a Mind: Statistics, Structure, and Abstraction*. Science, vol. 331, pages 1279 – 1285, 2011.

[Tong 2022] Z. Tong, Y. Song, J. Wang and L. Wang. *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*. ArXiv, vol. abs/2203.12602, 2022.

[Topol 2025] P. R. . E. J. Topol. *The Robot Doctor Will See You Now*. The New York Times, 2025.

[Tran 2014] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani and M. Paluri. *Learning Spatiotemporal Features with 3D Convolutional Networks*. 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2014.

[Tuyls 2020] K. Tuyls, S. Omidshafiei, P. Muller, Z. Wang, J. T. Connor, D. Hennes, I. Graham, W. Spearman, T. Waskett, D. Steele, P. Luc,

A. Recasens, A. Galashov, G. Thornton, R. Élie, P. Sprechmann, P. Moreno, K. Cao, M. Garnelo, P. Dutta, M. Valko, N. M. O. Heess, A. Bridgland, J. Pérolat, B. D. Vylder, A. Eslami, M. Rowland, A. Jaegle, R. Munos, T. Back, R. Ahamed, S. Bouton, N. Beauguerlange, J. Broshear, T. Graepel and D. Hassabis. *Game Plan: What AI can do for Football, and What Football can do for AI*. J. Artif. Intell. Res., vol. 71, pages 41–88, 2020.

[Twinanda 2015] A. P. Twinanda, E. O. Alkan, A. Gangi, M. de Mathelin and N. Padoy. *Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms*. International Journal of Computer Assisted Radiology and Surgery, vol. 10, pages 737–747, 2015.

[Twinanda 2016] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin and N. Padoy. *EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos*. IEEE Transactions on Medical Imaging, vol. 36, pages 86–97, 2016.

[Twinanda 2017] A. P. Twinanda. *Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos*. PhD thesis, Universite de Strasbourg, 2017. Thèse de doctorat dirigée par Mathelin, Michel de Image et vision Strasbourg 2017.

[Undre 2006] S. Undre, N. Sevdalis, A. N. Healey, S. A. W. Darzi and C. A. Vincent. *Teamwork in the operating theatre: cohesion or confusion?* Journal of evaluation in clinical practice, vol. 12 2, pages 182–9, 2006.

[van den Oord 2018] A. van den Oord, Y. Li and O. Vinyals. *Representation Learning with Contrastive Predictive Coding*. ArXiv, vol. abs/1807.03748, 2018.

[van der Maaten 2008] L. van der Maaten and G. E. Hinton. *Visualizing Data using t-SNE*. Journal of Machine Learning Research, vol. 9, pages 2579–2605, 2008.

[Vaswani 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin. *Attention Is All You Need*. NeurIPS, vol. abs/1706.03762, 2017.

[Verellen 1999] D. Verellen and F. Vanhavere. *Risk assessment of radiation-induced malignancies based on whole-body equivalent dose estimates for IMRT treatment in the head and neck region.* Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology, vol. 53 3, pages 199–203, 1999.

[Wagner 2024] L. Wagner, S. Jourdan, L. Mayer, C. Müller, L. Bernhard, S. Kolb, F. Harb, A. Jell, M. Berlet, H. Feussner, P. Buxmann, A. Knoll and D. Wilhelm. *Robotic scrub nurse to anticipate surgical instruments based on real-time laparoscopic video analysis*. Communications Medicine, vol. 4, page 156, 2024.

[Wang 2018] X. Wang and A. K. Gupta. *Videos as Space-Time Region Graphs*. ECCV, vol. abs/1806.01810, 2018.

[Wang 2020] J. Wang, J. Jiao and Y. Liu. *Self-supervised Video Representation Learning by Pace Prediction*. ECCV, vol. abs/2008.05861, 2020.

[Wang 2023a] X. Wang, R. Girdhar, S. X. Yu and I. Misra. *Cut and learn for unsupervised object detection and instance segmentation*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3124–3134, 2023.

[Wang 2023b] Y. Wang, X. Shen, Y. Yuan, Y. Du, M. Li, S. X. Hu, J. L. Crowley and D. Vaufreydaz. *Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut*. IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 12, pages 15790–15801, 2023.

[Wang 2023c] Z. Wang, P. Velickovic, D. Hennes, N. Tomaev, L. Prince, M. Kaisers, Y. Bachrach, R. Élie, W. K. Li, F. Piccinini, W. Spearman, I. Graham, J. T. Connor, Y. Yang, A. Recasens, M. Khan, N. Beauguerlange, P. Sprechmann, P. Moreno, N. M. O. Heess, M. Bowling, D. Hassabis and K. Tuyls. *TacticAI: an AI assistant for football tactics*. Nature Communications, vol. 15, 2023.

[Wee 2020] I. J. Y. Wee, L.-J. Kuo and J. C.-Y. Ngu. *A systematic review of the true benefit of robotic surgery: Ergonomics*. The International Journal of Medical Robotics and Computer Assisted Surgery, vol. 16, 2020.

[Wei 2022] C. Wei, H. Fan, S. Xie, C. Wu, A. L. Yuille and C. Feichtenhofer. *Masked Feature Prediction for Self-Supervised Visual Pre-Training*. CVPR, 2022.

[Weinland 2007] D. Weinland, E. Boyer and R. Ronfard. *Action Recognition from Arbitrary Views using 3D Exemplars*. 2007 IEEE 11th International Conference on Computer Vision, pages 1–7, 2007.

[Witmer 2022] H. D. D. Witmer, A. Dhiman, A. D. Jones, A. M. Laffan, D. Adelman and K. K. Turaga. *A Systematic Review of Operative Team Familiarity on Metrics of Efficiency, Patient Outcomes, Cost, and Team Satisfaction*. Annals of Surgery, vol. 276, pages e674 – e681, 2022.

[Wu 2016] Z. Wu, Y. Fu, Y.-G. Jiang and L. Sigal. *Harnessing Object and Scene Semantics for Large-Scale Video Understanding*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3112–3121, 2016.

[Wu 2021] C.-Y. Wu and P. Krahenbuhl. *Towards long-form video understanding*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1884–1894, 2021.

[Xie 2024] Y. Xie, K. Yang, N. Yang, W. Deng, X. Dai, T. Gu, Y. Wang, X. An, Y. Zhao, Z. Feng and J. Deng. *Croc: Pretraining Large Multimodal Models with Cross-Modal Comprehension*. ArXiv, vol. abs/2410.14332, 2024.

[Xu 2019] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie and Y. Zhuang. *Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10326–10335, 2019.

[Xu 2022] Y. Xu, J. Zhang, Q. Zhang and D. Tao. *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 38571–38584. Curran Associates, Inc., 2022.

[Yang 2017] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel*et al. Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy*, 2017.

[Yengera 2018] G. Yengera, D. Mutter, J. Marescaux and N. Padoy. *Less is More: Surgical Phase Recognition with Less Annotations through Self-Supervised Pre-training of CNN-LSTM Networks*. ArXiv, vol. abs/1805.08569, 2018.

[Zamudio 2023] J. Zamudio, F. F. Kanji, C. Lusk, D. Shouhed, B. R. Sanchez, K. Catchpole, J. T. Anger and T. N. Cohen. *Identifying workflow disruptions in robotic-assisted bariatric surgery: elucidating challenges experienced by surgical teams*. Obesity Surgery, vol. 33, no. 7, pages 2083–2089, 2023.

[Zappella 2013] L. Zappella, B. B. Haro, G. Hager and R. Vidal. *Surgical gesture classification from video and kinematic data*. Medical image analysis, vol. 17 7, pages 732–45, 2013.

[Zhang 2016] R. Zhang, P. Isola and A. A. Efros. *Colorful Image Colorization*. In European Conference on Computer Vision, 2016.

[Zhang 2021] Y. Zhang, I. Marsic and R. S. Burd. *Real-time medical phase recognition using long-term video understanding and progress gate method*. Medical Image Analysis, vol. 74, page 102224, 2021.

[Zhang 2024] H. Zhang, M. C. Leong, L. Li and W. Lin. *PeVL: Pose-Enhanced Vision-Language Model for Fine-Grained Human Action Recognition*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18857–18867, 2024.

[Zhou 2023] X. Zhou, A. Arnab, C. Sun and C. Schmid. *How can objects help action recognition?* 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2353–2362, 2023.

[Zhu 2020] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai. *Deformable DETR: Deformable Transformers for End-to-End Object Detection*. ArXiv, vol. abs/2010.04159, 2020.

[Zisimopoulos 2018] O. Zisimopoulos, E. Flouty, I. Luengo, P. Giataganas, J. Nehme, A. Chow and D. Stoyanov. *DeepPhase: Surgical Phase Recognition in CATARACTS Videos*. MICCAI, 2018.

[Zohar 2020] M. Zohar, O. Bar, D. Neimark, G. D. Hager and D. Asselmann. *Accurate detection of out of body segments in surgical video using semi-supervised learning*. In Medical Imaging with Deep Learning, pages 923–936. PMLR, 2020.

[Zou 2024] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao and Y. J. Lee. *Segment everything everywhere all at once*. Advances in Neural Information Processing Systems, vol. 36, 2024.

[Zuckerman 2024] I. Zuckerman, N. Werner, J. Kouchly, E. Huston, S. DiMarco, P. D. Dimusto and S. Laufer. *Depth over RGB: automatic evaluation of open surgery skills using depth camera*. International Journal of Computer Assisted Radiology and Surgery, vol. 19, pages 1349 – 1357, 2024.

[Özsoy 2023] E. Özsoy, T. Czempiel, E. P. Örnek, U. Eck, F. Tombari and N. Navab. *Holistic OR domain modeling: a semantic scene graph approach*. IJCARS, 2023.

Idris HAMOUD

Data-efficient Multimodal Learning by exploiting Scene Semantics for Operating Room Workflow Monitoring

Résumé :
L'analyse holistique de vidéos du bloc opératoire est essentielle pour le développement de modèles d'intelligence artificielle capables de reconnaître automatiquement et précisément les différentes étapes du workflow chirurgical. Cette reconnaissance automatique permettrait de créer des modèles d'aide à la décision améliorant la sécurité, l'efficacité et le temps d'utilisation du bloc opératoire. Les méthodes actuelles reposent sur l'apprentissage supervisé, nécessitant beaucoup de données étiquetées, et ne permettant pas un transfert facile dans des blocs operatoires disposant d'un positionnement de camera different. Cette thèse propose de nouvelles approches auto-supervisées pour développer des méthodes d'analyse du déroulement des activités opératoires, en mettant l'accent sur des modalités *abstraites* ou sémantiques telles que la detection d'objet ou l'estimation de pose des cliniciens. S'appuyant sur les avancées récentes de l'apprentissage auto-supervisé en vision par ordinateur, les méthodes proposées utilisent des autoencodeurs masqués, l'apprentissage contrastif multimodal et des tâches prétextuelles soigneusement conçues. L'utilisation de ces modalités moins coûteuses en annotations permettra la mise en place de ces méthodes dans des contextes cliniques réels.

Mots-Clés: Apprentissage profond - Vision par Ordinateur - Apprentissage auto-supervisé - Apprentissage Multimodal - Workflow au Bloc Opératoire - Analyse de Vidéo

Abstract :
Video recordings of operating room (OR) workflows are invaluable for studying and improving teamwork among clinicians. Automating the recognition of clinical activities in these videos is critical for applications such as modeling interactions and enhancing safety and operational efficiency. However, current methods largely depend on fully supervised training, making datasets even harder to generate and often failing to generalize across ORs with different camera setups. Existing self-supervised techniques focus on appearance-based tasks, overlooking vital semantic information like object detection and human pose data. Incorporating these semantic elements can narrow domain gaps and reduce the need for extensive labeling. This thesis proposes new self-supervised approaches to develop recognition approaches for monitoring OR workflows by emphasizing these "abstract" or semantic modalities. Such modalities are more cost-effective and easier to obtain than manual annotations. Building on recent advancements in self-supervised learning for computer vision, the proposed methods utilize masked autoencoders, multimodal contrastive learning, and carefully designed pretext tasks. Ultimately, this work aims to minimize labeling requirements and bolster the scalability and adaptability of surgical workflow monitoring.

Keywords: Deep learning - Computer Vision - Self-supervised Learning - Multimodal Learning - OR Workflow - Video Understanding